2018

Approximate Bayesian approaches and semiparametric methods for handling missing data

Hejian Sang
Iowa State University

Approximate Bayesian approaches and semiparametric methods for handling missing data

by

Hejian Sang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Jae Kwang Kim, Major Professor
Wayne A. Fuller
Vivekananda Roy
Cindy Yu
Zhengyuan Zhu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

I would like to dedicate this thesis to my parents. Without their support, I would not have been able to complete this work. I also would like to thank all my friends and family for their encouragement through my Ph.D life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my major professor Jae Kwang Kim for the patient guidance in various aspect of conducting research and the writing of scientific papers. I would like to thank my girl friend for her support. Moreover, I would like to appreciate committee members: Wayne A. Fuller, Vivekananda Roy, Cindy Yu and Zhengyuan Zhu, for their helpful advice and comments in preliminary exam.

I owe a huge thanks to Dr. Gyuhyeong Goh in Kansas State University and Dr. Kosuke Morikawa in the University of Tokyo. They provided me much assistance to complete this work.

I would like to acknowledge the faculty in the Center for Survey Statistics and Methodology (CSSM) at ISU. I am especially appreciated for the opportunity to work as a research assistant for CSSM.

ABSTRACT

This thesis consists of four research papers focusing on estimation and inference in missing data. In the first paper (Chapter 2), an approximate Bayesian approach is developed to handle unit nonresponse with parametric model assumptions on the response probability, but without model assumptions for the outcome variable. The proposed Bayesian method is also extended to incorporate the auxiliary information from full sample. In second paper (Chapter 3), a new Bayesian method using the Spike-and-Slab prior is proposed to handle the sparse propensity score estimation. The proposed method is not based on any model assumption on the outcome variable and is computationally efficient. In third paper (Chapter 4), we develop a robust semiparametric method based on the profile likelihood obtained from semiparametric response model. The proposed method uses the observed regression model and the semiparametric response model to achieve robustness. An efficient algorithm using fractional imputation is developed. The bootstrap testing procedure is also proposed to test ignorability assumption. In last paper (Chapter 5), we propose a novel semiparametric fractional imputation method using Gaussian mixture model for handling multivariate missingness. The proposed method is computationally efficient and leads to robust estimation. The proposed method is further extended to incorporate the categorical auxiliary information. Asymptotic properties are developed for each proposed methods. Both simulation studies and real data applications are conducted to check the performance of the proposed methods in this thesis.

CHAPTER 1. OVERVIEW

Missing data is frequently encountered in many areas of statistics. Ignoring missing data can lead to a biased estimation. The missing mechanism can mainly be categorized into three types. If the missing mechanism does not depend on data, it is missing completely at random (MCAR). Under MCAR, analysis methods only using complete data are consistent. However, MCAR is very limited in practice. The second missing mechanism is missing at random (Rubin, 1976) in the sense that missingness does not depend on missing values and only depends on observed data. MAR is a common assumption due to its simplicity. Under MAR, one of the popular methods of handling missing data is to build a model for the response mechanism and use the inverse of the estimated response probability to construct weights for estimating parameters. Such weighting method is called propensity score weighting (Rosenbaum and Rubin, 1983). The last missing mechanism is not missing at random and also referred as nonignorable missingness, when missingness also depends on unobserved values. NMAR is more challenging than MAR, since the response model cannot be estimated without extra assumptions.

In the first paper, we are interested in developing Bayesian inference for propensity score estimation. One of the main advantages of Bayesian inference is that all the uncertainty in the estimation process can be built into the Bayesian computation automatically. That is, there is no need to conduct variance estimation separately in the Bayesian inference. While the Bayesian method is widely used in many areas of statistics, the literature on the Bayesian approach of propensity score estimation is sparse. In the first paper, we propose a novel approach featuring approximate Bayesian computation based on the summary statistics (Beaumont et al., 2002).

However, when the dimension of the covariates for the propensity score is large, the full response model including all the covariates may have several problems. While spare model is wildly used in the linear regression to improve efficiency, the sparsity effect on the propensity score estimation is

somehow unclear. To the best of our knowledge, not much work has been done for sparse propensity score estimation in the missing data context. In second paper, we propose a Bayesian approach for the sparse propensity score estimation. Our main goal is to develop a valid inference procedure for estimating equations with the sparse propensity score adjustment. One of the greatest advantages of the Bayesian approach is that both estimating the parameter of interest and eliminating irrelevant covariates can be simultaneously performed in the posterior inference. To introduce the sparse posterior distribution, we propose to use stochastic search variable selection with the Spike-and-Slab prior. The proposed Bayesian method is implemented by data augmentation algorithm (Tanner and Wong, 1987; Wei and Tanner, 1990).

In addition to MAR, we develop a semiparametric estimation using profile likelihood and test for handing NMAR in our third paper. Under nonignorable nonresponse, we believe that response variable plays a critical role in the response model. The generalized linearity assumption of response in the response model can be limited. The proposed method uses the generalized partially linear model with a nonparametric function of response. The estimation method is developed from maximizing the profile likelihood. An efficient computation algorithm is proposed based on the fractional imputation (Kim, 2011). Furthermore, we propose a hypothesis test to check if the response mechanism is missing at random. A bootstrap method is proposed to compute the empirical distribution of the test statistic. The proposed method is robust, since the observed regression model can be justified from the data directly and the response mechanism is a flexible function of response.

Our last paper focuses on handling multivariate missingness. For multivariate missing data with arbitrary missing patterns, imputation methods are developed to preserve the correlation structure in the imputed data. Conditional models for the different missing patterns calculated directly from the observed patterns may not be compatible with each other. The parametric fractional imputation used the joint distribution to create imputed values, but correct specification of the joint model is challenging under missing data. Furthermore, valid inference after multiple imputation requires congeniality and self-efficiency (Meng, 1994), which is not necessary satisfied in many practical

problems (Kim et al., 2006; Yang and Kim, 2016b). Fractional imputation does not suffer such problems. Note that parametric imputation requires correct model specification. Nonparametric imputation methods, such as kernel regression imputation (Cheng, 1994; Wang and Chen, 2009), are robust but may be subject to curse of dimensionality. It is important to develop a unified, robust and efficient imputation method. The proposed semiparametric method fills in this important gap by considering a flexible method for imputation. In this paper, to achieve robustness against model misspecification, we develop an imputation procedure based on Gaussian mixture models (GMM). GMM is a very flexible model that can be used to handle outliers, heterogeneity and skewness. It is semiparametric in the sense that the number of mixture component is chosen automatically from the data. The computation is relatively simple and efficient.

The rest of this thesis is organized as follows. In Chapter 2, we introduce our proposed approximate Bayesian inference on propensity score method. In Chapter 3, we present Bayesian sparse propensity score estimation for unit nonresponse approach. A profile likelihood approach to semiparametric estimation with nonignorable nonresponse is shown in Chapter 4. In Chapter 5, we propose a semiparametric fractional imputation method using Gaussian mixture models for handling multivariate missing data. Some summary and remarks are presented in Chapter 6.

Bibliography

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162(4):2025–2035.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):81–87.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1):119–132.

Kim, J. K., Michael Brick, J., Fuller, W. A., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):509–521.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 37(1):490–517.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704.

Yang, S. and Kim, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika*, 103(1):244–251.

CHAPTER 2. AN APPROXIMATE BAYESIAN INFERENCE USING PROPENSITY SCORE ESTIMATION UNDER UNIT NONRESPONSE

Hejian Sang Jae Kwang Kim

Abstract

Nonresponse weighting adjustment using the response propensity score is a popular tool for handling unit nonresponse. Statistical inference after the nonresponse weighting adjustment is complicated because the effect of estimating the propensity model parameter needs to be incorporated into finding inference. In this paper, we propose an approximate Bayesian approach to handle unit nonresponse with parametric model assumptions on the response probability, but without model assumptions for the outcome variable. The proposed Bayesian method is calibrated to the frequentist inference in that the credible region obtained from the posterior distribution asymptotically matches to the frequentist confidence interval obtained from the Taylor linearization method. The proposed Bayesian method is also extended to incorporate the auxiliary information from full sample. Results from limited simulation studies confirm the validity of the proposed methods. The proposed method is applied to data from a Korean longitudinal survey.

key words: Approximate Bayesian computation; Missing at random; Nonresponse weighting adjustment.

2.1 Introduction

Missing data is frequently encountered in many areas of statistics. When the response mechanism is missing at random in the sense of Rubin (1976), one of the popular methods of handling missing data is to build a model for the response probability and use the inverse of the estimated response probability to construct weights for estimating parameters. Such weighting method is often called propensity score weighting and the resulting estimator is called propensity score estimator

(Rosenbaum and Rubin, 1983). The propensity score method has been well studied in the literature. For examples, see Rosenbaum (1987), Flanders and Greenland (1991), Robins et al. (1994), Robins et al. (1995), and Kim and Kim (2007). However, all the above researches were developed under the frequentist approaches. Variance estimates using Taylor linearization or bootstrap are used to make frequentist inferences.

In this paper, we are interested in developing Bayesian inference for propensity score estimation. One of the main advantages of Bayesian inference is that all the uncertainty in the estimation process can be built into the Bayesian computation automatically. That is, there is no need to develop variance estimation separately in the Bayesian inference. While the Bayesian method is widely used in many areas of statistics, the literature on the Bayesian approach of propensity score estimation is sparse. An (2010) proposed a Bayesian propensity score estimator with jointly modeling the response mechanism and the outcome variable. Specifying correct outcome model is difficult under missing data and incorrect specification may lead to biased inference. McCandless et al. (2009) and Kaplan and Chen (2012) also assumed joint models and obtained Bayesian credible regions in the context of casual inference.

In this paper, we propose a new Bayesian approach of propensity score estimation without making any model assumptions on the outcome variable. Since no parametric model assumptions on the outcome variable are used, there is no explicit likelihood function corresponding to θ, the main parameter of interest, which makes the problem difficult to solve.

To overcome such challenges, we develop a novel Bayesian approach using the idea of approximate Bayesian computation (ABC) based on the summary statistics (Beaumont et al., 2002). In the proposed method, the sampling distribution of summary statistics, which is the estimating equation itself, can be used to replace the likelihood part in deriving the posterior distribution. See Sunnåker et al. (2013), Toni et al. (2009), Csilléry et al. (2010) and Soubeyrand and Haon-Lasportes (2015) for examples. It is also similar in spirit to Bayesian generalized method of moments of Yin et al. (2009). In the proposed Bayesian method, the credible region obtained from the posterior distribution asymptotically matches the frequentist confidence interval obtained from the Taylor

linearization method. The computation for the proposed method is relatively simple and easy to understand.

Note that, the propensity score estimation does not use full sample information, in the sense that the propensity score estimator of the auxiliary variables is not necessary equal to the full sample estimator. To incorporate this additional auxiliary information, the optimal propensity score estimation using augmented estimation equations is developed. See Zhou and Kim (2012), Cao et al. (2009), and Imai and Ratkovic (2014). We extend the proposed Bayesian propensity score estimation method to obtain the optimal Bayesian propensity estimator by including additional propensity score estimation of the auxiliary variables.

The rest of the paper is organized as follows. In §2.2, we introduce the basic setup of the general propensity score estimation problem. The proposed method is presented in §2.3. The main result and asymptotic theory are discussed in §2.4. In §2.5, we developed a related method by extending our proposed method to incorporate the auxiliary information observed throughout the sample. The finite sample performance of the proposed methods is examined in an extensive simulation study in §2.6. An application of the proposed methods to a longitudinal survey is presented in §2.7. Some concluding remarks and future work are discussed in §2.8. The proofs and technique derivations are given in Appendix.

2.2 Basic Setup

Suppose that we are interested in estimating θ defined through $E\{U(\theta; X, Y)\} = 0$ for some estimating function $U(\theta; X, Y)$. Let $(x_i, y_i), i = 1, \cdots, n$, be independently and identically distributed realizations of random variable (X, Y). Under complete response, we can obtain a consistent estimator of θ by solving

$$\frac{1}{n}\sum_{i=1}^{n} U(\theta; x_i, y_i) = 0 \tag{2.1}$$

for θ without model assumption on Y. We assume that the solution to (2.1) is unique almost everywhere to avoid the model non-identifiability issue.

Now, suppose that the auxiliary variable X is always observed and the response variable Y is subject to missingness. In this case, we can define the response indicator function for unit i as

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

We assume the response mechanism is missing at random in the sense of Rubin (1976). Furthermore, assume that δ_i are independently generated from a Bernoulli distribution with

$$\text{pr}(\delta_i = 1 \mid x_i, y_i) = \pi(\phi; x_i) \tag{2.2}$$

for some unknown parameter vector ϕ and $\pi(\cdot)$ is a known link function.

When nonresponse exists, we cannot apply (2.1) directly. Instead, using the response probability in (2.2), we can obtain the propensity score estimator of θ by the following two steps:

Step 1. Compute the maximum likelihood estimator $\hat{\phi}$ of ϕ by maximizing

$$L_1(\phi) = \prod_{i=1}^{n} \pi(\phi; x_i)^{\delta_i} \{1 - \pi(\phi; x_i)\}^{1-\delta_i}. \tag{2.3}$$

Step 2. Compute the propensity score estimator of θ, say $\hat{\theta}_{PS}$, by solving

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\hat{\phi}; x_i)} U(\theta; x_i, y_i) = 0.$$

Under the above setup, we propose a new Bayesian approach to make inference from the posterior distribution. An advantage of the Bayesian approach is that we can incorporate the uncertainty in estimating ϕ into the Bayesian computation automatically. Furthermore, prior information about ϕ or θ can be naturally handled in the Bayesian framework.

2.3 Proposed Method

We now present the proposed Bayesian method in the case of missing at random. Let $X_n = (x_1, x_2, \cdots, x_n)$, $\Delta_n = (\delta_1, \delta_2, \cdots, \delta_n)$ and Y_{obs} denote the observed part of $Y_n = (y_1, y_2, \cdots, y_n)$. Under the Bayesian framework, the posterior distribution $p(\phi, \theta \mid X_n, \Delta_n, Y_{obs})$ can be obtained by

$$p(\phi, \theta \mid X_n, \Delta_n, Y_{obs}) = \frac{L(\phi, \theta \mid X_n, \Delta_n, Y_{obs})\pi(\phi)\pi(\theta)}{\int L(\phi, \theta \mid X_n, \Delta_n, Y_{obs})\pi(\phi)\pi(\theta)\text{d}\phi\text{d}\theta}, \tag{2.4}$$

where $L(\phi, \theta \mid X_n, \Delta_n, Y_{obs})$ is the joint likelihood function of (ϕ, θ) based on (X_n, Δ_n, Y_{obs}) and $\{\pi(\phi), \pi(\theta)\}$ are prior distributions for ϕ and θ. Unfortunately, the likelihood function for θ is not available.

In the approximate Bayesian method, we approximate the likelihood part by the sampling distribution of the summary statistics. In the context of propensity score estimation, the summary statistics for (ϕ, θ) is the estimating function itself. That is, $\{S(\phi), U_{PS}(\theta, \phi)\}$ is the summary statistics for (ϕ, θ), where

$$S(\phi) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi(\phi; x_i)} - \frac{1-\delta_i}{1-\pi(\phi; x_i)}\right\}\frac{\partial \pi(\phi; x_i)}{\partial \phi} =: \frac{1}{n}\sum_{i=1}^{n} s(\phi; x_i, \delta_i) \tag{2.5}$$

and

$$U_{PS}(\theta, \phi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\phi; x_i)} U(\theta; x_i, y_i). \tag{2.6}$$

Thus, we can use

$$\hat{p}(\phi, \theta \mid X_n, \Delta_n, Y_{obs}) = \frac{g\{S(\phi), U_{PS}(\theta, \phi) \mid \phi, \theta\}\pi(\phi)\pi(\theta)}{\int g\{S(\phi), U_{PS}(\theta, \phi) \mid \phi, \theta\}\pi(\phi)\pi(\theta)\mathrm{d}\phi\mathrm{d}\theta}, \tag{2.7}$$

as an approximation for the posterior distribution in (2.4), where $g\{S(\phi), U_{PS}(\theta, \phi) \mid \phi, \theta\}$ is the sampling distribution of $S(\phi)$ and $U_{PS}(\theta, \phi)$.

To obtain the sampling distribution, under certain regularity conditions, we can establish the asymptotic distribution of $\{S(\phi), U_{PS}(\theta, \phi)\}$ as

$$\sqrt{n}\left\{S^T(\phi), U_{PS}^T(\theta, \phi)\right\}^T \to N(0, \Sigma)$$

in distribution, where

$$\Sigma = \Sigma(\phi, \theta) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

is a positive-definite matrix. Therefore, the sampling distribution $g\{S(\phi), U_{PS}(\theta, \phi) \mid \phi, \theta\}$ is approximated by a normal distribution with mean 0 and variance $n^{-1}\Sigma$.

Now, since we can decompose the joint likelihood function as $L(\theta, \phi \mid X_n, \Delta_n, Y_{obs}) = L_1(\phi \mid X_n, \Delta_n) L_2(\theta \mid X_n, \Delta_n, Y_{obs}, \phi)$, we can avoid the approximate Bayesian technique in generating ϕ

and only apply it in generating θ. Thus, the following two-step method can be used in generating (ϕ, θ) from the approximate posterior distribution:

Step 1. Generate ϕ^* from

$$\frac{L_1(\phi \mid X_n, \Delta_n)\pi(\phi)}{\int L_1(\phi \mid X_n, \Delta_n)\pi(\phi)\mathrm{d}\phi},$$

where $L_1(\phi \mid X_n, \Delta_n)$ is defined in (2.3).

Step 2. Generate θ^* from

$$p(\theta \mid X_n, \Delta_n, Y_{obs}, \phi^*) = \frac{\hat{p}(\theta, \phi^* \mid X_n, \Delta_n, Y_{obs})}{\hat{p}(\phi^* \mid X_n, \Delta_n, Y_{obs})}, \tag{2.8}$$

where $\hat{p}(\theta, \phi \mid X_n, \Delta_n, Y_{obs})$ is defined in (2.7) and

$$\hat{p}(\phi \mid X_n, \Delta_n, Y_{obs}) = \frac{g_1\{S(\phi) \mid \phi\}\pi(\phi)}{\int g_1\{S(\phi) \mid \phi\}\pi(\phi)\mathrm{d}\phi}.$$

Using (2.7) and (2.8), the posterior distribution in (2.8) reduces to

$$p(\theta \mid X_n, \Delta_n, Y_{obs}, \phi^*) \propto \frac{g\{S(\phi^*), U_{PS}(\theta, \phi^*) \mid \phi^*, \theta\}\pi(\theta)}{g_1\{S(\phi^*) \mid \phi^*\}},$$

which yields to

$$p(\theta \mid X_n, \Delta_n, Y_{obs}, \phi^*) = \frac{g_2\{U_{PS}(\theta, \phi^*) \mid S(\phi^*), \theta\}\pi(\theta)}{\int g_2\{U_{PS}(\theta, \phi^*) \mid S(\phi^*), \theta\}\pi(\theta)\mathrm{d}\theta},$$

where $g_2\{U_{PS}(\theta, \phi) \mid S(\phi), \theta\}$ is the density function of the conditional distribution of $U_{PS}(\theta, \phi)$ given $S(\phi)$. Thus, we can simplify *Step 2* as follows:

Step 2. Given ϕ^*, generate θ^* from

$$\theta^* \sim p(\theta \mid X_n, \Delta, Y_{obs}, \phi^*) \propto g_2\{U_{PS}(\theta, \phi^*) \mid S(\phi^*), \theta\}\pi(\theta). \tag{2.9}$$

$g_2\{U_{PS}(\theta, \phi) \mid S(\phi), \theta\}$ is the normal density function with mean $\kappa S(\phi)$ and variance $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, where $\kappa = \Sigma_{21}\Sigma_{11}^{-1}$. To generate θ^* from (2.9), we use a consistent estimator of Σ in the sampling distribution g_2.

To summarize, the proposed Bayesian propensity score method can be described as follows:

Step 1. Generate ϕ^* from

$$\phi^* \sim p(\phi \mid X_n, \Delta_n) \propto L_1(\phi \mid X_n, \Delta_n)\pi(\phi).$$

Step 2. Given ϕ^*, generate θ^* from

$$\theta^* \sim p(\theta \mid X_n, \delta_n, Y_{obs}, \phi^*) \propto \hat{g}_2\left\{U_{PS}(\theta, \phi^*) \mid S(\phi^*), \theta\right\}\pi(\theta),$$

where $\hat{g}_2\left\{(\cdot \mid S(\phi^*), \theta\right\}$ is the estimated density function of $g_2\left\{\cdot \mid S(\phi^*), \theta\right\}$ with Σ replaced by $\hat{\Sigma}(\phi^*, \theta)$. See Appendix 2.9 for details.

2.4 Asymptotic Properties

To formulate the asymptotic properties of the proposed Bayesian propensity method, denote $\zeta = (\phi, \theta)$ and $\zeta_0 = (\phi_0, \theta_0)$, where ζ_0 is the true parameter value generating the sample. Let the joint propensity score estimating equations be $H_n(\zeta) = \{S(\phi), U_{PS}(\phi, \theta)\}$. The asymptotic properties of the posterior distribution include posterior consistency and posterior asymptotic normality.

To establish the consistency of the parameter estimate and the interval estimate under the frequentist propensity score estimation, we assume the following regularity conditions:

Assumption 1. As n goes to infinity, $H_n(\zeta) \to \eta(\zeta)$ in probability uniformly, where $\eta(\zeta) = E\left\{H_n(\zeta)\right\}$. That is, $\sup_\zeta \|H_n(\zeta) - \eta(\zeta)\| \to 0$ in probability.

Assumption 2. The mapping $\zeta \mapsto H_n(\zeta)$ is continuous and has exactly one zero $\hat{\zeta}_n$ almost everywhere.

Assumption 3. There exists an unique ζ_0 such that $\inf_{\zeta: d(\zeta, \zeta_0) \geq \epsilon} \|\eta(\zeta)\| > 0 = \eta(\zeta_0)$, for any $\epsilon > 0$, where d is a distance function.

Assumption 4. There exists a neighbor of ζ_0, denoted by $N_n(\zeta_0)$, on which with probability one all $H_n(\zeta)$ are continuously differentiable and the Jacobian $\partial H_n(\zeta)/\partial\zeta$ converges uniformly to a non-stochastic and non-singular limit. Here, $N_n(\zeta_0)$ is a ball with center ζ_0 and radius r_n, where r_n satisfies $r_n \to 0$ and $r_n\sqrt{n} \to \infty$.

Assumption 5. For any $\zeta \in N_n(\zeta_0)$, $H_n(\zeta)$ is Lipschitz continuous for ζ and $E\{H_n^{\otimes 2}(\zeta_0)\} < \infty$, where $A^{\otimes 2} = AA^T$.

Assumptions 1–5 are the standard conditions to achieve the consistency of the propensity score estimation and asymptotic normality, in the sense of

$$\sqrt{n}\left(\hat{\zeta}_n - \zeta_0\right) \to N\{0, W(\zeta_0)\} \tag{2.10}$$

in distribution, where $W(\zeta_0) = A(\zeta_0)^{-1}\Sigma(\zeta_0)A^T(\zeta_0)^{-1}$, $A(\zeta) = \partial\eta(\zeta)/\partial\zeta^T$ and $\Sigma(\zeta_0) = E\{H_n^{\otimes 2}(\zeta_0)\}$.

We now make additional assumptions to establish the posterior consistency and convergence in distribution:

Assumption 6. The prior distribution $\pi(\zeta)$ is absolutely continuous in $N_n(\zeta_0)$ and has a positive density on ζ_0 .

Assumption 7. For $\zeta \in N_n(\zeta_0)$, the variance estimator is consistent, in the sense of $\hat{\Sigma}(\zeta) = \Sigma(\zeta)\{1 + o_p(1)\}$.

Assumption 6 is a common assumption for the prior and the flat prior satisfies this condition. The positive support on $\boldsymbol{\zeta}_0$ ensures the posterior distribution covers the true value. Assumption 7 is the sufficient condition for approximating the posterior distribution in *Step 2* of the proposed Bayesian propensity score method.

Theorem 2.1. *Under assumptions 1–7, the posterior distribution $p(\zeta \mid X_n, \Delta_n, Y_{obs})$, generated from the proposed Bayesian propensity score method in §2.3, satisfies*

$$\|p\{\sqrt{n}(\zeta - \zeta_0)|X_n, \Delta_n, Y_{obs}\} - g\{\sqrt{n}(\zeta - \zeta_0); 0, W(\zeta_0)\}\| \to 0 \tag{2.11}$$

in probability and

$$\text{pr}\left\{\lim_{n\to\infty}\int_{N_n(\zeta_0)} p(\zeta|X_n, \Delta_n, Y_{obs})\text{d}\zeta = 1\right\} = 1, \tag{2.12}$$

where $g\{\cdot; 0, W(\zeta_0)\}$ is the density of the approximate normal distribution in (2.10).

The proof is shown in Appendix 2.10. Result (2.11) is the convergence of the posterior distribution to normality and result (2.12) is the strong posterior consistency. By (2.11), the confidence region using the proposed Bayesian method is asymptotically equivalent to the frequentist confidence

region based on asymptotic normality of $\hat{\zeta}_n$. Thus, our proposed Bayesian method is calibrated to frequentist inference using asymptotic normality of $\hat{\theta}_{PS}$.

2.5 Optimal Estimation

We now extend the proposed method to incorporate additional information from the full sample. Note that the propensity score estimator applied to $\mu_x = E(X)$ can be computed as the solution to

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi(\hat{\phi}; x_i)}(x_i - \mu_x) = 0$$

which is not necessarily equal to $\hat{\mu}_x = n^{-1}\sum_{i=1}^{n} x_i$. Including this extra information in the propensity score estimation, if done properly, will improve the efficiency of the resulting propensity score estimator. In the frequentist propensity score method, incorporating such extra information can be implemented by generalized method of moments and it is sometimes called the optimal propensity score estimation.

To include such extra information, we can add

$$\begin{aligned} U_{PS,x}(\mu_x, \phi) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi(\phi; x_i)}(x_i - \mu_x) \\ U_x(\mu_x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x) \end{aligned}$$

in addition to the original propensity score estimating equations (2.5) and (2.6).

To formally describe the proposed Bayesian method, define $\psi = (\theta, \mu_x)$ and

$$U_J(\psi, \theta) = \{U_{PS}(\theta, \phi), U_{PS,x}(\mu_x, \phi), U_x(\mu_x)\}.$$

The joint likelihood function of (ϕ, ψ) can be decomposed as

$$L(\phi, \psi \mid X_n, \Delta_n, Y_{obs}) = L_1(\phi \mid X_n, \Delta_n, Y_{obs}) L_2(\psi \mid X_n, \Delta_n, Y_{obs}, \phi). \tag{2.13}$$

Similarly to §2.3, $L_2(\psi \mid X_n, \Delta_n, Y_{obs}, \phi)$ is not well defined without any model assumptions on X and Y. From (2.13), $L_1(\phi \mid X_n, \Delta_n, Y_{obs})$ can be used to generate ϕ^*. To generate the posterior

draw ψ^*, we can use, similarly to (2.8), the two approximate distributions to derive the conditional distribution of ψ given ϕ^* as follows:

$$\psi^* \sim p(\psi \mid X_n, \Delta_n, Y_{obs}, \phi^*) = \frac{\hat{p}(\psi, \phi \mid X_n, \Delta_n, Y_{obs})}{\hat{p}(\phi \mid X_n, \Delta_n, Y_{obs})},$$

where $\hat{p}(\psi, \phi \mid X_n, \Delta_n, Y_{obs}) \propto g\left\{U_J(\psi, \phi), S(\phi) \mid \phi, \psi\right\} \pi(\phi)\pi(\psi)$, $\hat{p}(\phi \mid X_n, \Delta_n, Y_{obs}) \propto g_1\left\{S(\phi) \mid \phi\right\} \pi(\phi)$, and $g(\cdot \mid \phi, \psi)$ can be approximated by the asymptotic normal distribution from $\sqrt{n}\left\{S^T(\phi), U_J^T(\psi, \phi)\right\}^T \to N(0, \Sigma)$. Therefore,

$$p(\psi \mid X_n, \Delta_n, Y_{obs}, \phi) \propto \frac{g\left\{U_J(\psi, \phi), S(\phi) \mid \phi, \psi\right\} \pi(\psi)}{g_1\left\{S(\phi) \mid \phi\right\}} = g_2\left\{U_J(\psi, \phi) \mid S(\phi), \psi\right\} \pi(\psi),$$

where $g_2\left\{U_J(\psi, \phi) \mid S(\phi), \psi\right\}$ is the conditional density function.

Thus, the implementation of the proposed optimal Bayesian propensity score method can be described as the following two steps:

Step 1. Generate ϕ^* from

$$\phi^* \sim p(\phi \mid X_n, \Delta_n) = \frac{L_1(\phi \mid X_n, \Delta_n)\pi(\phi)}{\int L_1(\phi \mid X_n, \Delta_n)\pi(\phi)\mathrm{d}\phi}.$$

Step 2. Given ϕ^*, generate ψ^* from

$$\psi^* \sim p(\psi \mid X_n, \Delta_n, Y_{obs}, \phi^*) \propto g_2\left\{U_J(\psi, \phi^*) \mid S(\phi^*), \psi\right\} \pi(\psi). \tag{2.14}$$

The posterior distribution in (2.14) can be obtained by Metropolis–Hastings algorithm. The proposed optimal Bayesian propensity method incorporates the full sample information and calibrates to the frequentist optimal estimation using the generalized method of moments.

2.6 Simulation Study

We perform a limited simulation study to validate our proposed methods and to check the effect of prior distributions. The performance of the proposed Bayesian methods with informative priors and non-informative priors is compared with the frequentist propensity score method. The simulation study is a 2×2 factorial design, where the factors are outcome regression models for $E(y \mid x)$ and sample size.

For the outcome regression models, we consider the following two candidates:

$$M_1: \quad y = \beta_0 + \beta_1 x + e \qquad (\beta_0, \beta_1) = (1, 1)$$
$$M_2: \quad y = \beta_0 + \beta_1 x^2 + e \qquad (\beta_0, \beta_1) = (1, 0.5)$$

where the error distribution is $e \sim N(0, 0.25)$. The superpopulation models M_1 and M_2 are used to check the performance of the proposed methods under the linear and nonlinear models. The explanatory variable x is generated from $N(1, 1)$ independently.

For the response mechanism, the response indicator function δ_i are independently generated from a Bernoulli distribution with probability

$$p_i(\phi_0, \phi_1) = \frac{\exp(\phi_0 + \phi_1 x_i)}{1 + \exp(\phi_0 + \phi_1 x_i)} \tag{2.15}$$

with $(\phi_0, \phi_1) = (0.1, 1)$, which makes the overall response rate approximately equal to 70%.

For each setup, we generate random samples of size $n = 50$ or 500 independently with $B = 2,000$ replications. From each realized sample, we specify a logistic regression model in (2.15) as the response model. For each Monte Carlo sample, we use the following methods to make inference for $\theta = E(Y)$:

1. PS: Frequentist propensity score approach based on Taylor linearization. The point estimator $(\hat{\theta}_{PS}, \hat{\phi})$ is computed from

$$U_{PS}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\phi; x_i)} (y_i - \theta) = 0$$
$$S(\phi) = \frac{1}{n} \sum_{i=1}^{n} \{\delta_i - \pi(\phi; x_i)\}(1, x_i)^T = 0.$$

   The confidence intervals are constructed by $\hat{\theta}_{PS} \pm 1{\cdot}96\sqrt{\hat{V}_{PS}}$, where $\hat{V}_{PS} = \widehat{\text{var}}(\hat{\theta}_{PS})$ is obtained by the Taylor linearization.

2. Bayesian PS (BPS): The proposed Bayesian method based on the parametric model assumption in (2.15). For prior specifications, we consider the following four cases:

   a: $\pi(\phi) \propto 1$ and $\pi(\theta) \propto 1$.

b: $\pi(\phi) \sim N(b_0, B_0)$ and $\pi(\theta) \propto 1$.

c: $\pi(\phi) \propto 1$ and $\pi(\theta) \sim N(\mu_0, s_0)$.

d: $\pi(\phi) \sim N(b_0, B_0)$ and $\pi(\theta) \sim N(\mu_0, s_0)$,

where b_0 is the true value of (ϕ_0, ϕ_1), $B_0 = \text{diag}(1, 1)$, μ_0 is the true value of $E(Y)$ and $s_0 = 1$. The estimators for (ϕ, θ) are obtained by the mean of the draws from the approximate posterior distribution. The credible intervals can be constructed by quantiles of the posterior distribution. Under the non-formative prior a, the proposed Bayesian method should calibrate to propensity score method asymptotically. For prior b, where prior of ϕ is informative, but prior of θ is non-informative, we explore the effect of prior of ϕ on estimating of θ. For prior c, we use the non-informative prior for ϕ and the informative prior of θ to check the effect of prior of θ. The prior d is to check the effect of jointly informative priors.

3. Optimal PS (OPS): Use the generalized method of moments as

$$C_n(\phi, \theta, \mu_x) = \frac{1}{n} \sum_{i=1}^{n} \begin{Bmatrix} \{\delta_i - \pi(\phi; x_i)\} (1, x_i)^T \\ \delta_i \pi(\phi; x_i)^{-1} (y_i - \theta) \\ \delta_i \pi(\phi; x_i)^{-1} (x_i - \mu_x) \\ x_i - \mu_x \end{Bmatrix}.$$

The OPS estimator is obtained by minimizing $C_n^T W^{-1} C_n$, where $W = \text{var}(C_n)$. See §5.4 of Kim and Shao (2013) for details.

4. OBPS: Optimal Bayesian PS method discussed in §2.5. The prior distributions $\pi(\theta, \phi) \propto 1$ and $\pi(\mu_x) \propto 1$. The credible intervals can be constructed quantiles of the posterior distributions.

For each of the four methods, 95% confidence intervals for θ are computed from Monte Carlo samples. The simulation result is presented in Table 2.1.

From Table 2.1, where $n = 500$ and the population model is linear (M_1), we can see that, overall, the proposed BPS achieves the same standard errors and the coverage probabilities with the

Table 2.1: Simulation results from a Monte Carlo study of size $B = 2,000$. : "bias" is the Monte Carlo bias. "std" is the Monte Carlo standard error. "AL" is the average length of the confidence (credible) intervals. "CP" is the coverage probability for the corresponding confidence (credible) interval. "PS" is the propensity score estimation. "BPS_a", "BPS_b","BPS_c" and "BPS_d" are the Bayesian propensity score method with prior a,b,c and d, respectively. "OPS" is the optimal propensity score estimation. "OBPS" is the optimal Bayesian propensity score method with non-informative priors.

n	method	$M1$				$M2$			
		bias	std	AL	CP	bias	std	AL	CP
500	PS	0.00	0.06	0.22	0.95	-0.00	0.06	0.25	0.95
	BPS_ a	-0.00	0.06	0.22	0.95	-0.00	0.06	0.25	0.95
	BPS_b	-0.00	0.06	0.22	0.95	-0.00	0.06	0.25	0.95
	BPS_c	-0.00	0.06	0.22	0.95	-0.00	0.06	0.25	0.95
	BPS_d	-0.00	0.06	0.22	0.95	-0.00	0.06	0.25	0.95
	OPS	0.00	0.05	0.20	0.95	-0.01	0.06	0.24	0.95
	OBPS	0.00	0.05	0.20	0.95	-0.01	0.06	0.24	0.95
50	PS	-0.01	0.18	0.67	0.94	-0.01	0.20	0.75	0.93
	BPS_a	-0.07	0.21	0.75	0.94	-0.03	0.20	0.76	0.92
	BPS_b	-0.05	0.19	0.69	0.93	-0.02	0.20	0.75	0.92
	BPS_c	-0.03	0.19	0.64	0.93	-0.02	0.19	0.72	0.92
	BPS_d	-0.01	0.18	0.63	0.93	-0.03	0.19	0.75	0.93
	OPS	0.01	0.17	0.61	0.93	-0.02	0.20	0.75	0.93
	OBPS	0.01	0.17	0.63	0.94	-0.02	0.20	0.75	0.93

frequentist PS method regardless of whether priors are informative or flat. This is consistent with Theorem 2.1 in the sense that the posterior distribution converges to the asymptotic distribution of maximum likelihood estimator as the sample size becomes large enough. Also we find that the proposed OBPS method is calibrated to the OPS method with showing the same performance in term of standard errors and length of credible (confidence) intervals. Comparing four methods, the OBPS and OPS always perform better than BPS and PS methods with incorporating full sample information. When the population model is quadratic in Table 2.1, the same conclusions of M_1 can be obtained. When the outcome regression model is quadratic (M_2), the proposed two Bayesian methods also obtain the same performance with the frequentist methods.

To explore the effect of priors in the PS estimation, we also set the sample small size as $n = 50$. From Table 2.1, the proposed BPS method with flat priors obtains larger standard errors and wider

credible intervals than the frequentist PS method, which yields to better or equivalent coverage probabilities. Under the BPS method, the informative priors of ϕ and θ help to reduce variability and bias. The prior information of θ only (prior c) achieves the better perfomrance than prior b, where we use only informative prior for ϕ. Also, the BPS with jointly informative priors is better than the PS method in term of narrower confidence length. Comparing the OBPS and OPS, we can see that the proposed OBPS method provides similar credible intervals with better coverage probabilities than the OPS. In summary, the proposed Bayesian methods outperform the frequentist methods under the small sample size.

2.7 Application

In this section, we apply the proposed Bayesian propensity score methods to Korea Labor and Income Panel Survey data. A brief description of the panel survey can be found at http://www.kli.re.kr/klips/en/about/introduce.jsp. The study variable (y) is the average monthly income for the current year and the auxiliary variable (x) can be demographic variables, such as the age groups and sex. Let (x_i, y_{it}) be the observations for household i in panel year t. The KLIPS has $n = 5,013$ households and $T = 8$ panel years. We treat the first panel observations as the baseline measurements, and there are no missing data in the first year. In the panel survey, x_i are completely observed and y_{it} are subject to missingness, for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$. Let δ_{it} be the response indicator function of y_{it}. Define

$$\delta_{it} = \begin{cases} 1 & \text{if we observe } y_{it} \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in estimating the probability of full response

$$\pi_i = \text{pr}(\delta_{i1} = 1, \cdots, \delta_{iT} = 1 \mid x_i, y_{i,obs}), \tag{2.16}$$

where $y_{i,obs} = (y_{i1}, \cdots, y_{iT})$ represent the observed responses for household i. The inverse of the π_i in (2.16) can be used as the propensity weight for the penal survey. For monotone missing data, in the sense of $\delta_{it} = 1$ implying $\delta_{i,t-1} = 1, \cdots, \delta_{i1} = 1$, the probability reduces to $\pi_i = \pi_{i1}\pi_{i2}\cdots\pi_{iT}$, where $\pi_{it} = \text{pr}(\delta_{it} = 1 \mid \delta_{i,t-1} = 1, x_i, y_{i1}, \cdots, y_{i,t-1})$ under missing at random assumption.

For arbitrary missing patterns, we first define $\delta_{it}^* = \prod_{k=1}^{t} \delta_{ik}$. Note that $\delta_{it}^* = 1$ implies that $\delta_{i,t-1}^* = 1$. Furthermore,

$$\begin{aligned}\text{pr}(\delta_{i1} = 1, \cdots, \delta_{iT} = 1 \mid x_i, y_{i,obs}) &= \prod_{k=2}^{T} \text{pr}(\delta_{ik}^* = 1 \mid \delta_{i,k-1}^* = 1, x_i, y_{i,k-1}) \\ &= \prod_{k=2}^{T} \text{pr}(\delta_{ik} = 1 \mid \delta_{i,k-1}^* = 1, x_i, y_{i,k-1}) \\ &= \pi_{i2}\pi_{i3}\cdots\pi_{iT} = \pi_i,\end{aligned}$$

where $\pi_{i1} = 1$ for all samples.

Thus, we can build a parametric model for $\pi_{it} = \text{pr}(\delta_{it} = 1 \mid \delta_{i,t-1}^* = 1, x_i, y_{i,t-1})$ and estimate the parameters sequentially. Instead of using the frequentist approach of Zhou and Kim (2012), we apply the Bayesian propensity method in §2.3 and the optimal Bayesian method in §2.5 to incorporate the extra information in x.

We are interested in estimating the average income for the final year and constructing confidence intervals for the parameters. Assume the response mechanism follows

$$\pi(\phi_t; x_i, y_{i,t-1}) = \text{pr}(\delta_{it} = 1 \mid \delta_{i,t-1}^* = 1, x_i, y_{i,t-1}) = \frac{1}{1 + \exp\{-(x_i^T, y_{i,t-1})\phi_t\}}, \tag{2.17}$$

which is known up to parameter ϕ_t. Thus, we allow that the response probability at year t depends on the last year income y_{t-1}, but not on the current year income. Assume δ_{it}, given $\delta_{i,t-1}^* = 1, x_i$, and $y_{i,t-1}$, independently follow Bernoulli distribution with probability $\pi(\phi_t; x_i, y_{i,t-1})$ in (2.17). Therefore, we can apply the proposed Bayesian propensity method sequentially for each t. Then the joint estimating equations are $U_n(\phi_2, \phi_3, \cdots, \phi_T, \theta) = 0$, where

$$U_n(\phi_2, \phi_3, \cdots, \phi_T, \theta) = n^{-1} \sum_{i=1}^{n} \pi_i^{-1}(\delta_{iT}^* y_{iT} - \theta) \tag{2.18}$$

and $\theta = E(Y_T)$. The proposed Bayesian propensity method can be applied to obtain the posterior distribution of $(\phi_2, \cdots, \phi_T, \theta)$ with known likelihood function of ϕ_t and approximated sampling distribution of $U_n(\phi_2, \phi_3, \cdots, \phi_T, \theta)$.

To improve the efficiency of the point estimator, we also apply the optimal Bayesian propensity method to the same sample. In addition to equations in (2.18), we add $\sum_{i=1}^{n} \delta_{iT}^* \pi_i^{-1}(x_i - \mu_x) = 0$

and $\sum_{i=1}^{n}(x_i - \mu_x) = 0$, where μ_x is the marginal proportion vector for demographical covariates. Therefore, the posterior distribution of θ can be obtained by applying the proposed algorithm in §5. For a comparison, we also considered a naive method which does not use the propensity model and apply the Bayesian method in the complete cases (CC) only. The numerical results are presented below.

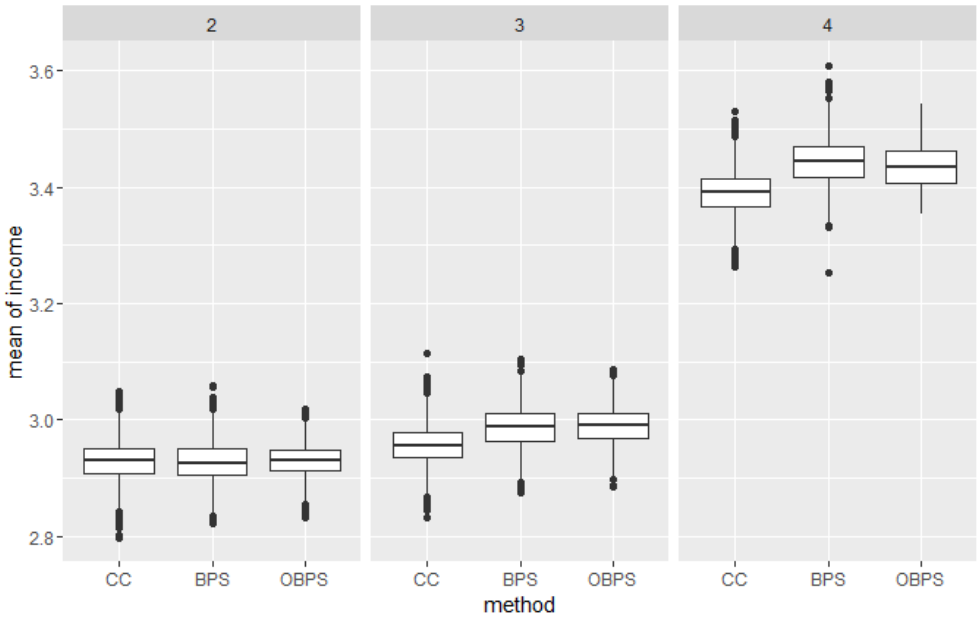


Figure 2.1: Boxplots for posterior distribution of θ (Magnitude 1,000,000 Won) by different methods and panels $T = 2, 3, 4$. "CC" denotes the Bayesian method only using the complete data. "BPS" is the proposed Bayesian propensity score method. "OBPS" is the optimal Bayesian propensity method with incorporating information of X.

From Figure 2.1, all three methods provide similar estimators for the average income θ.The trend of average income goes up as year T increases. For year $T = 2$, all three methods provide similar mean estimates. But the OBPS method is the most efficient. For year $T = 3$, we see that the CC method provides lower mean estimate than BPS or OBPS, which is due to the nonresponse bias in the CC method. This phenomenon becomes more obvious for year $T = 4$. Also, the lengths of confidence intervals increase as T increases, since the fully observed sample size is decreasing due to panel attrition. The CC method presents smaller values of θ for $T = 4$, which suggests

more panel attrition for higher income households. Both BPS and OBPS provide similar mean estimates. But the OBPS method has narrower confidence intervals, which confirms the efficiency of the OBPS method.

2.8 Concluding Remarks

A new Bayesian inference using propensity score method is developed using the idea of Approximate Bayesian computation. The proposed method can be widely applicable due to popularity of propensity score method. The proposed Bayesian approach is calibrated to frequentist inference in the sense that the proposed method provides the same inferential results with its frequentist version asymptotically (Little, 2012). The calibration property holds if the sample size is large enough. If the prior is informative then the resulting Bayesian inference could be more efficient than frequentist inference due to its natural incorporation of the prior information. Thus, the proposed method is applicable when combining information from different sources.

Causal inference, including estimation of average treatment effect from observational studies, can be one promising application area of the propensity score method (Morgan and Winship, 2014; Hudgens and Halloran, 2008). Developing tools for causal inference using the Bayesian propensity score method will be an important extension of this research. Also, Bayesian model selection method (Ishwaran and Rao, 2005) can be naturally applied to this setup. Such extensions will be topics for future research.

Appendix includes a brief description about the consistent variance estimator of $\{S(\phi), U_{PS}(\theta,\phi)\}$ in *Step 2* and the proof of Theorem 1 in §4.

2.9 Appendix A: The consistent variance estimator in step 2

Note that,

$$\text{var}\left\{\sqrt{n}S(\phi)^T, \sqrt{n}U_{PS}^T(\theta,\phi)\right\}^T \to \Sigma(\phi,\theta),$$

in probability. Since $\{(x_1, y_1\delta_1, \delta_1), \cdots (x_n, y_n\delta_n, \delta_n)\}$ are independent, the consistent estimator of $\Sigma(\phi,\theta)$ is

$$\hat{\Sigma}(\phi,\theta) = \frac{1}{n}\sum_{i=1}^{n}\begin{Bmatrix} s(\phi; x_i, \delta_i) \\ \delta_i \pi^{-1}(\phi; x_i) U(\theta; x_i, y_i) \end{Bmatrix}^{\otimes 2},$$

where $A^{\otimes 2} = AA^T$ and $s(\phi; x, \delta)$ is the score function of ϕ.

2.10 Appendix B: Proof of Theorem 1

First, we can decompose the posterior distribution as

$$p\left\{\sqrt{n}(\zeta-\zeta_0) \mid X_n, \Delta_n, Y_{obs}\right\} = \quad p\left\{\sqrt{n}(\phi-\phi_0) \mid X_n, \Delta_n, Y_{obs}\right\} p\left\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\}.$$

From the asymptotic distribution of $(\hat{\phi}, \hat{\theta})$, we have

$$\sqrt{n}\begin{pmatrix}\hat{\phi}-\phi_0 \\ \hat{\theta}-\theta_0\end{pmatrix} \to N(0, W) \tag{2.19}$$

in distribution, where

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22}\end{pmatrix}.$$

Thus, given $\sqrt{n}(\phi-\phi_0)$, we have

$$\sqrt{n}(\theta-\theta_0) \mid \phi \to N(W_{12}W_{11}^{-1}\sqrt{n}(\phi-\phi_0), W_{22} - W_{21}W_{11}^{-1}W_{12})$$

in distribution. Note that, the distribution is conditional on ϕ, which is equivalent to giving $\sqrt{n}(\phi - \phi_0)$. Denote $\mu(\phi) = W_{12}W_{11}^{-1}\sqrt{n}(\phi - \phi_0)$ and $W_{22\cdot 1} = W_{22} - W_{21}W_{11}^{-1}W_{12}$. Similarly, we can decompose the asymptotic distribution of (2.19) as

$$g\left\{\sqrt{n}(\zeta - \zeta_0); 0, W\right\} = g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} g\left\{\sqrt{n}(\theta - \theta_0); \mu(\phi), W_{22\cdot 1}\right\},$$

where $g(\cdot; \mu, S)$ is the normal density function with mean μ and variance S.

Note that, the propose Bayesian method uses the explicit likelihood of ϕ and the approximate distribution of θ. Thus, we can obtain that

$$\begin{aligned}
&\| p\left\{\sqrt{n}(\zeta - \zeta_0) \mid X_n, \Delta_n, Y_{obs}\right\} - g\left\{\sqrt{n}(\zeta - \zeta_0); 0, W\right\} \| \\
&= \| p\left\{\sqrt{n}(\phi - \phi_0) \mid X_n, \Delta_n, Y_{obs}\right\} p\left\{\sqrt{n}(\theta - \theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \\
&- g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} g\left\{\sqrt{n}(\theta - \theta_0); \mu(\phi), W_{22\cdot 1}\right\} \| \\
&= \| p\left\{\sqrt{n}(\phi - \phi_0) \mid X_n, \Delta_n, Y_{obs}\right\} p\left\{\sqrt{n}(\theta - \theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \\
&- g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} p\left\{\sqrt{n}(\theta - \theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \\
&+ g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} p\left\{\sqrt{n}(\theta - \theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \\
&- g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} g\left\{\sqrt{n}(\theta - \theta_0); \mu(\phi), W_{22\cdot 1}\right\} \|.
\end{aligned}$$

Using the triangular inequality, it is sufficient to show that

$$\begin{aligned}
&\| p\left\{\sqrt{n}(\zeta - \zeta_0) \mid X_n, \Delta_n, Y_{obs}\right\} - g\left\{\sqrt{n}(\zeta - \zeta_0); 0, W\right\} \| \\
&\leq \| p\left\{\sqrt{n}(\phi - \phi_0) \mid X_n, \Delta_n, Y_{obs}\right\} p\left\{\sqrt{n}(\theta - \theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \\
&- g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} p\left\{\sqrt{n}(\theta - \theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \| \\
&+ \| g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} p\left\{\sqrt{n}(\theta - \theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \\
&- g\left\{\sqrt{n}(\phi - \phi_0); 0, W_1\right\} g\left\{\sqrt{n}(\theta - \theta_0); \mu(\phi), W_{22\cdot 1}\right\} \| \to 0,
\end{aligned}$$

in probability. From *step 2*, $p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\} \propto \hat{g}_2\{U_{PS}(\theta;\phi) \mid S(\phi), \theta\}\pi(\theta)$ is uniformly bounded by c_1, when the posterior distribution is appropriate. Then, we can obtain that

$$\begin{aligned}
&\|p\{\sqrt{n}(\phi-\phi_0) \mid X_n, \Delta_n, Y_{obs}\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\}\| \\
&-g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\}\| \\
&\le c_1\|p\{\sqrt{n}(\phi-\phi_0) \mid X_n, \Delta_n, Y_{obs}\} - g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\|.
\end{aligned}$$

step 1 is a standard Bayesian method. By Bernstein-von Mises theorem (Van der Vaart, 1998, Chapter 10), we have the following conclusion:

$$\|p\{\sqrt{n}(\phi-\phi_0) \mid X_n, \Delta_n, Y_{obs}\} - g\left(\sqrt{n}(\phi-\phi_0); 0, W_1\right)\| \to 0$$

in probability, which yields to

$$\begin{aligned}
&\|p\{\sqrt{n}(\phi-\phi_0) \mid X_n, \Delta_n, Y_{obs}\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\}\| \\
&-g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\}\| \to 0
\end{aligned} \tag{2.20}$$

in probability.

Then, next step is to show

$$\begin{aligned}
&\|g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\} \\
&-g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\, g\{\sqrt{n}(\theta-\theta_0); \mu(\phi), W_{22\cdot 1}\}\| \to 0
\end{aligned} \tag{2.21}$$

in probability. We can rewrite (2.21) as

$$\begin{aligned}
&\|g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\} \\
&-g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\, g\{\sqrt{n}(\theta-\theta_0); \mu(\phi), W_{22\cdot 1}\}\| \\
&\le \|g\{\sqrt{n}(\phi-\phi_0); 0, W_1\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\} \\
&-g\{\sqrt{n}S(\phi); 0, \Sigma_{11}\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\}\| \\
&+\|g\{\sqrt{n}S(\phi); 0, \Sigma_{11}\}\, p\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\} \\
&-g\{\sqrt{n}H_n(\zeta); 0, \Sigma\}\| \\
&= J_1 + J_2.
\end{aligned}$$

Thus, it is sufficient to show that $J_1 \to 0$ and $J_2 \to 0$ in probability. For the first claim $J_1 \to 0$ in probability, we can conclude it from

$$\begin{aligned} J_1 &= \| g\left\{\sqrt{n}(\phi-\phi_0); 0, W_1\right\} p\left\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \\ &- g\left\{\sqrt{n}S(\phi); 0, \Sigma_{11}\right\} p\left\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} \| \\ &\leq c_1 \| g\left\{\sqrt{n}(\phi-\phi_0); 0, W_1\right\} - g\left\{\sqrt{n}S(\phi); 0, \Sigma_{11}\right\} \| \to 0 \end{aligned} \tag{2.22}$$

in probability, where the convergence in probability holds because $\sqrt{n}(\hat{\phi}-\phi_0) \to N(0, W_1)$ is asymptotically equivalent to $\sqrt{n}S(\phi) \to N(0, \Sigma_{11})$. Using the extended dominated convergence theorem, we can show the convergence in probability holds.

Note that, *step 2* is to generate θ^* from

$$\theta^* \sim p(\theta \mid X_n, \delta_n, Y_{obs}, \phi^*) \propto \hat{g}_2\left\{U_{PS}(\theta, \phi^*) \mid S(\phi^*), \theta\right\} \pi(\theta)$$

Therefore, we can show that

$$\begin{aligned} &\| g\left\{\sqrt{n}S(\phi); 0, \Sigma_{11}\right\} p\left\{\sqrt{n}(\theta-\theta_0) \mid X_n, \Delta_n, Y_{obs}, \phi\right\} - g\left\{\sqrt{n}H_n(\zeta); 0, \Sigma\right\} \| \\ &= \| g\left\{\sqrt{n}S(\phi); 0, \Sigma_{11}\right\} c(\phi)\hat{g}_2\left\{\sqrt{n}U_{PS}(\theta, \phi) \mid S(\phi), \theta\right\} \pi(\theta) - g\left\{\sqrt{n}H_n(\zeta); 0, \Sigma\right\} \|, \\ &= \| g\left\{\sqrt{n}S(\phi); 0, \Sigma_{11}\right\} c(\phi)\hat{g}_2\left\{\sqrt{n}U_{PS}(\theta, \phi) \mid S(\phi), \theta\right\} \pi(\theta) \\ &- g\left\{\sqrt{n}S(\phi); 0, \Sigma_{11}\right\} g\left\{\sqrt{n}U_{PS}(\theta, \phi); \Sigma_{21}\Sigma_{11}^{-1}S(\phi), \Sigma_{22\cdot 1}\right\} \| \\ &\leq c_2 \| c(\phi)\hat{g}_2\left\{\sqrt{n}U_{PS}(\theta, \phi) \mid S(\phi), \theta\right\} \pi(\theta) - g\left\{\sqrt{n}U_{PS}(\theta, \phi); \Sigma_{21}\Sigma_{11}^{-1}S(\phi), \Sigma_{22\cdot 1}\right\} \|, \end{aligned}$$

where $c(\phi)$ is the normalized constant, $g\left(\cdot; 0, \Sigma_{11}\right)$ is bounded by c_2 and $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

By the assumption (7), we can have $\hat{V}_U = \Sigma_{22\cdot 1}\left\{1 + o_p(1)\right\}$ and $\hat{\kappa} = \Sigma_{21}\Sigma_{11}^{-1}\left\{1 + o_p(1)\right\}$. Then, we can derive that

$$\begin{aligned} &c(\phi)\hat{g}_2\left\{\sqrt{n}U_{PS}(\theta, \phi) \mid S(\phi), \theta\right\} \pi(\theta) - g\left\{\sqrt{n}U_{PS}(\theta, \phi); \Sigma_{21}\Sigma_{11}^{-1}S(\phi), \Sigma_{22\cdot 1}\right\} \\ &= c(\phi) g\left\{\sqrt{n}U_{PS}(\theta, \phi); \hat{\kappa}S(\phi), \hat{V}_U\right\} \pi(\theta) - g\left\{\sqrt{n}U_{PS}(\theta, \theta); \Sigma_{21}\Sigma_{11}^{-1}S(\phi), \Sigma_{22\cdot 1}\right\} \\ &= \frac{\exp\left[-0.5n\left\{U_{PS}(\theta, \phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}^T \Sigma_{22\cdot 1}^{-1}\left\{U_{PS}(\theta, \phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}\left\{1 + o_p(1)\right\}\right] \pi(\theta)}{\int \exp\left[-0.5n\left\{U_{PS}(\theta, \phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}^T \Sigma_{22\cdot 1}^{-1}\left\{U_{PS}(\theta, \phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}\left\{1 + o_p(1)\right\}\right] \pi(\theta) d\theta} \\ &- g\left\{\sqrt{n}U_{PS}(\theta, \phi); \Sigma_{21}\Sigma_{11}^{-1}S(\phi), \Sigma_{22\cdot 1}\right\} \end{aligned}$$

Note that, when θ_1 solves $U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi) = 0$, we have the following conclusion:

$$\exp\left[-0.5n\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}^T \Sigma_{22\cdot1}^{-1}\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}\{1+o_p(1)\}\right]\pi(\theta) = \pi(\theta_1).$$

If $U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi) \neq 0$, then

$$\exp\left[-0.5n\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}^T \Sigma_{22\cdot1}^{-1}\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}\{1+o_p(1)\}\right]$$
$$\to 0,$$

in probability. Therefore, we can show that the approximate integration of the conditional distribution goes to the following point mass:

$$\int \exp\left[-0.5n\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}^T \Sigma_{22\cdot1}^{-1}\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}\{1+o_p(1)\}\right]\pi(\theta)d\theta$$
$$\to |2\pi\Sigma_{22\cdot1}|^{-1/2}\pi(\theta_1).$$

Thus, we have

$$= \frac{\exp\left[-0.5n\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}^T \Sigma_{22\cdot1}^{-1}\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}\{1+o_p(1)\}\right]\pi(\theta)}{\int \exp\left[-0.5n\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}^T \Sigma_{22\cdot1}^{-1}\left\{U_{PS}(\theta,\phi) - \Sigma_{21}\Sigma_{11}^{-1}S(\phi)\right\}\{1+o_p(1)\}\right]\pi(\theta)d\theta}$$
$$-g\left\{\sqrt{n}U_{PS}(\theta,\phi); \Sigma_{21}\Sigma_{11}^{-1}S(\phi), \Sigma_{22\cdot1}\right\} \to 0,$$

in probability. By the extended dominated convergence theorem, we can show

$$\|c(\phi)\hat{g}_2\left\{\sqrt{n}U_{PS}(\theta,\phi) \mid S(\phi),\theta\right\}\pi(\theta) - g\left\{\sqrt{n}U_{PS}(\theta,\phi); \Sigma_{21}\Sigma_{11}^{-1}S(\phi), \Sigma_{22\cdot1}\right\}\| \to 0, \tag{2.23}$$

in probability.

(2.22) and (2.23) completes the proof of (2.21). Combining (2.20) and (2.21), we have

$$\|p\left\{\sqrt{n}(\zeta-\zeta_0)|X_n,\Delta_n,Y_{obs}\right\} - g\left\{\sqrt{n}(\zeta-\zeta_0); 0, W(\zeta_0)\right\}\| \to 0, \tag{2.24}$$

in probability.

Next, we are going to show the consistency of the posterior distribution. From the asymptotic distribution (2.19), we can define

$$C_{n,\alpha} = \left\{\zeta : n(\zeta-\zeta_0)^T W^{-1}(\zeta-\zeta_0) \leq \chi_p^2(\alpha)\right\},$$

where the $\chi_p^2(\alpha)$ is the α quantile of the Chi-square distribution with p degrees of freedom.

Furthermore, from a property of the Raylei Quotient (Horn and Johnson, 1990), there exists a matrix O such that

$$OW^{-1}O^T = \text{diagonal}\left\{\lambda_1, \cdots, \lambda_p\right\},$$

where $OO^T = I_p$ and $0 < \lambda_1 \leq \lambda_2, \cdots, \leq \lambda_p$. Thus we obtain

$$x^T \left(nW^{-1}\right) x \geq n\lambda_1 x^T x. \tag{2.25}$$

Applying the conclusion (2.25), we can obtain the following two inequalities:

$$\|\zeta - \zeta_0\| \leq \lambda_1^{-1/2}\sqrt{(\zeta - \zeta_0)^T nW^{-1}(\zeta - \zeta_0)} \leq \lambda_1^{-1/2}\sqrt{\chi_p^2(\alpha)/n}. \tag{2.26}$$

Next, from (2.26), we can conclude that

$$\lim_{n\to\infty} \text{pr}\left\{\|\zeta - \zeta_0\| \leq \lambda_1^{-1/2}\sqrt{\chi_p^2(\alpha)/n}\right\} \geq \alpha,$$

which leads to

$$\lim_{n\to\infty} \text{pr}\left\{\zeta \in C_{n,\alpha}, \|\zeta - \zeta_0\| \leq 2\lambda_1^{-1/2}\sqrt{\chi_p^2(\alpha)/n}\right\} \geq \alpha. \tag{2.27}$$

Since we have defined $N_n(\zeta_0)$ in a neighborhood with center ζ_0 and radius r_n, where r_n satisfies $r_n \to 0$ and $\sqrt{n}r_n \to \infty$. From (2.27),

$$\lim_{n\to\infty} \text{pr}\left\{\zeta \in C_{n,\alpha}, \|\zeta - \zeta_0\| \leq r_n\right\} \geq \alpha,$$
$$\lim_{n\to\infty} \text{pr}\left\{C_{n,\alpha} \subset N_n(\zeta_0)\right\} \geq \alpha.$$

Therefore,

$$\lim_{n\to\infty} \text{pr}\left\{\int_{N_n(\zeta_0)} g(\zeta; \zeta_0, n^{-1}W)d\zeta \geq \int_{C_{n,\alpha}} g(\zeta; \zeta_0, n^{-1}W)d\zeta\right\} \geq \alpha,$$

which is equivalent to

$$\lim_{n\to\infty} \text{pr}\left\{\int_{N_n(\zeta_0)} g(\zeta; \zeta_0, n^{-1}W)d\zeta \geq \alpha\right\} \geq \alpha. \tag{2.28}$$

Conclusion (2.28) holds for any $\alpha \in (0,1)$. Thus,

$$\text{pr}\left\{\lim_{n\to\infty}\int_{N_n(\zeta_0)} g(\zeta;\zeta_0,n^{-1}W)d\zeta = 1\right\} = 1. \tag{2.29}$$

Using the triangular inequality, we can obtain

$$\begin{aligned}\textstyle\int_{N_n(\zeta_0)} p(\zeta \mid X_n,\Delta_n,Y_{obs})d\zeta \quad &\geq \int_{N_n(\zeta_0)} g(\zeta;\zeta_0,n^{-1}W)d\zeta \\ &\quad - \int_{N_n(\zeta_0)} |p(\zeta \mid X_n,\Delta_n,Y_{obs}) - g(\zeta;\zeta_0,n^{-1}W)|d\zeta.\end{aligned}$$

From (2.29), we can show that the probability can be bounded by the following lower bound:

$$\begin{aligned}&\text{pr}\left\{\lim_{n\to\infty}\int_{N_n(\zeta_0)} p(\zeta \mid X_n,\Delta_n,Y_{obs})d\zeta\right. \\ &\left.\geq 1 - \lim_{n\to\infty}\int_{N_n(\zeta_0)} |p(\zeta \mid X_n,\Delta_n,Y_{obs}) - g(\zeta;\zeta_0,n^{-1}W)|d\zeta\right) = 1.\end{aligned} \tag{2.30}$$

From (2.24), we can obtain that, for any $\epsilon \in (0,1)$,

$$\text{pr}\left\{\lim_{n\to\infty}\int_{N_n(\zeta_0)} |p(\zeta \mid X_n,\Delta_n,Y_{obs}) - g(\zeta;\zeta_0,n^{-1}W)|d\zeta > \epsilon\right\} < \epsilon.$$

Thus, plugging into (2.30), we can obtain

$$\text{pr}\left\{\lim_{n\to\infty}\int_{N_n(\zeta_0)} p(\zeta \mid X_n,\Delta_n,Y_{obs})d\zeta \geq 1-\epsilon\right\} \geq 1-\epsilon,$$

for any $\epsilon \in (0,1)$. Therefore, we can conclude that,

$$\text{pr}\left\{\lim_{n\to\infty}\int_{N_n(\zeta_0)} p(\zeta \mid X_n,\Delta_n,Y_{obs})d\zeta\right\} = 1,$$

which completes the proof of Theorem 2.1.

Bibliography

An, W. (2010). Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociol. Methodol.*, 40(1):151–189.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162(4):2025–2035.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, 25(7):410–418.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statist. Med*, 10(5):739–747.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge university press.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *J. Am. Statist. Assoc.*, 103(482):832–842.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Statist. B.*, 76(1):243–263.

Ishwaran, H. and Rao, J. S. (2005). Spike and Slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.*, 33(2):730–773.

Kaplan, D. and Chen, J. (2012). A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, 77(3):581–609.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Can. J. Statist.*, 35(4):501–514.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press.

Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *J. Offic. Statist.*, 28(3):309–334.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statist. Med*, 28(1):94–112.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.*, 89(427):846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.*, 90(429):106–121.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *J. Am. Statist. Assoc.*, 82(398):387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Soubeyrand, S. and Haon-Lasportes, E. (2015). Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC. *Statist. Probab. Lett.*, 107:84–92.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS computational biology*, 9(1):e1002803.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Statist. Interface*, 6(31):187–202.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3. Cambridge university press.

Yin, G. et al. (2009). Bayesian generalized method of moments. *Bayesian Analysis*, 4(2):191–207.

Zhou, M. and Kim, J. K. (2012). An efficient method of estimation for longitudinal surveys with monotone missing data. *Biometrika*, 99(3):631–648.

CHAPTER 3. BAYESIAN SPARSE PROPENSITY SCORE ESTIMATION FOR UNIT NONRESPONSE

Hejian Sang Gyuhyeong Goh Jae Kwang Kim

Abstract

Nonresponse weighting adjustment using propensity score (PS) is a popular tool for handling unit nonresponse. However, including all the auxiliary variables into the propensity model can lead to inefficient estimation and the consistency is not guaranteed if the dimension of the covariates is large. In this paper, a new Bayesian method using the Spike-and-Slab prior is proposed to handle the sparse propensity score estimation. The proposed method is not based on any model assumption on the outcome variable and is computationally efficient. Instead of doing model selection and parameter estimation separately as in most frequentist methods, the proposed method simultaneously selects the true sparse response probability model and provides consistent parameter estimation and corresponding inference, which can be quite involved in the frequentist methods. The finite-sample performance of the proposed method is investigated in limited simulation studies, including a partially simulated real data example from the Korean Labor and Income Panel Survey.

key words: Approximate Bayesian computation, Data augmentation, Missing at random, Spike-and-Slab prior, Sparsity.

3.1 Introduction

Nonresponse in the collected data is a common problem in survey sampling, clinical trials, and many other areas of research. Ignoring nonresponse can lead to a biased estimation unless the missing mechanism is completely missing at random (Rubin, 1976). To handle nonresponse, various statistical methods have been developed. The propensity score weighting is one of the most popular tools for adjusting bias due to nonresponse, which builds on a model for the response probability

and uses the inverse of the estimated response probability as the weights for estimating parameters. Rosenbaum and Rubin (1983) showed that the propensity score adjustment is sufficient to remove the nonresponse bias under the correct response probability model. The propensity score weighting method is well established in the literature. See Rosenbaum (1987), Flanders and Greenland (1991), Robins et al. (1994), Robins et al. (1995) and Kim and Kim (2007). However, when the dimension of the covariates for the propensity score is large, the full response model including all the covariates may have several problems. First, the computation for the parameter estimation can be problematic as it involves high dimensional matrix inversion and the convergence is not guaranteed. Second, estimating zero coefficients in the propensity model increases the variability of the propensity scores and thus leads to inefficient estimates of the model parameters. Furthermore, the asymptotic normality of the PS estimator is not guaranteed if the dimension of the covariates is large. That is, the assumptions for the Central Limit Theorem (CLT) may not be satisfied if we include all the covariates into the propensity model. Therefore, a proper model selection to obtain a sparse propensity model is an important practical problem in the propensity score estimation.

Sparsity is a natural and important characteristic of statistical models. While sparsity is wildly used in the linear regression to improve efficiency, the sparsity effect on the propensity score estimation is somehow unclear. In the context of propensity score weighting, sparsity occurs when, among all the covariates under consideration, only a few of them are significantly involved in the true response mechanism. It is well known that traditional estimation methods such as the maximum likelihood estimation and the least squares estimation ignoring sparsity may yield poor estimates with large variance (Tibshirani, 1996; Zou and Hastie, 2005). Likewise, when sparsity is present in the propensity score, the propensity score estimation using the full model is less efficient than the method using the sparse model, even when the sample size is sufficiently large. See Lemma 3.1 in §3.2. There are many attempts to tackle sparse estimation in the classical regression problems. See Fan and Li (2001), Zou (2006), Park and Casella (2008), Kyung et al. (2010) for example. However, to the best of our knowledge, not much work has been done for sparse propensity score estimation in the missing data context.

In this paper, we propose a Bayesian approach for the sparse propensity score estimation. Our main goal is to develop a valid inference procedure for estimating equations with the sparse propensity score adjustment. One of the greatest advantages of the Bayesian approach is that both estimating the parameter of interest and eliminating irrelevant covariates can be simultaneously performed in the posterior inference. To introduce the sparse posterior distribution, we propose to use stochastic search variable selection with the Spike-and-Slab prior, which is a mixture of flat distribution and degenerate distribution at zero, or a mixture of their approximations (Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997; Narisetty et al., 2014). However, there is still a big challenge in implementing the Bayesian variable selection method in the propensity score (PS) estimation. In the estimating equations using the PS method, the likelihood function for the parameter of interest is unspecified. To resolve this issue, we derive an approximate likelihood from the sampling distribution of PS estimator. The proposed Bayesian method is implemented by data augmentation algorithm (Tanner and Wong, 1987; Wei and Tanner, 1990). The computation of posterior distribution is quite fast and efficient. The proposed method is justified using asymptotic theory and extensive simulation studies.

The rest of this paper is organized as follows. In §3.2, we introduce the basic setup of the PS estimation. The technical details of our proposal are described in §3.3. Model selection consistency and the asymptotic theory are established in §3.4. The performance of the proposed method is examined through simulation studies in §3.5. Some discussion is presented in §3.6. Proofs and derivations are given in Appendix.

3.2 Basic Setup

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n independent and identically distributed (IID) realizations from a random vector (X, Y), where Y is a scalar response and X is a p-dimensional vector of covariates. Suppose we are interested in estimating parameter $\theta \in \Theta$, which is the unique solution to the population estimating equation $E\{U(\theta; X, Y)\} = 0$. Under complete response, a consistent

estimator of θ can be obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n} U(\theta; x_i, y_i) = 0. \tag{3.1}$$

However, if nonresponse occurs, the estimating equation in (3.1) cannot be used directly.

To handle the missing data problem, the propensity score method using response propensity model can be used. To introduce the PS method, suppose that x_i are fully observed and y_i are subject to missingness. Let δ_i be the response indicator of y_i, that is,

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{if } y_i \text{ is missing.} \end{cases}$$

Assume that δ_i are independently distributed from a Bernoulli distribution with the success probability of $\Pr(\delta_i = 1|x_i, y_i)$. We further assume that the missing mechanism is missing at random (MAR) in the sense that

$$\Pr(\delta_i = 1|x_i, y_i) = \Pr(\delta_i = 1|x_i).$$

Following Rosenbaum and Rubin (1983), we define the propensity score for the i-th observation as

$$\Pr(\delta_i = 1|x_i) = \pi(\phi; x_i) = G\left(x_i^{\mathrm{T}}\phi\right), \tag{3.2}$$

where $G : \mathbb{R} \to [0, 1]$ is a known distribution function and $\phi = (\phi_1, \phi_2, \ldots, \phi_p)^{\mathrm{T}}$ is a p-dimensional unknown parameter. Then the propensity score estimator of θ, say $\hat{\theta}_{\mathrm{PS}}$, can be obtained by solving

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi(\hat{\phi}; x_i)} U(\theta; x_i, y_i) = 0, \tag{3.3}$$

with respect to θ, where $\hat{\phi}$ is a consistent estimator of ϕ. From the response model in (3.2), we can easily obtain the maximum likelihood estimator (MLE) of ϕ by maximizing the log-likelihood function,

$$l_n(\phi) = \sum_{i=1}^{n} \log f(\delta_i|x_i; \phi), \tag{3.4}$$

where

$$f(\delta_i|x_i; \phi) = \{\pi(x_i; \phi)\}^{\delta_i} \{1 - \pi(x_i; \phi)\}^{1-\delta_i}.$$

However, when ϕ is sparse, that is, ϕ contains many zero values, the MLE often involves large variance and fails to be consistent (Zou, 2006). Such phenomenon unfavorably leads to a poor inference about the parameter of interest θ. The following lemma illustrates the effect of including extra covariates in the PS estimation.

Lemma 3.1. *Suppose $X = (X_1, X_2)$ and the response mechanism is MAR. Let $\hat{\theta}_{PS}$ be the solution to*

$$\sum_{i=1}^{n} \frac{\delta_i}{Pr(\delta_i = 1|X_{1i}, X_{2i})} U(\theta; X_{1i}, y_i) = 0,$$

and $\hat{\theta}_{SPS}$ be the solution to

$$\sum_{i=1}^{n} \frac{\delta_i}{Pr(\delta_i = 1|X_{1i})} U(\theta; X_{1i}, y_i) = 0.$$

In this case, ignoring the smaller order terms, we have

$$E\left(\hat{\theta}_{PS}\right) = E\left(\hat{\theta}_{SPS}\right),$$
$$Var\left(\hat{\theta}_{SPS}\right) \leq Var\left(\hat{\theta}_{PS}\right).$$

Proof of Lemma 3.1 is presented in Appendix A. By Lemma 3.1, we can see that the propensity model including unnecessary covariates increases the variance of the resulting PS estimator. However, including important covariates into model is still critical to reduce the nonresponse bias.

Penalized likelihood estimation techniques have been proposed to overcome the drawback of MLE for high dimensional covariate problems. Similarly, we may achieve sparse and consistent estimation for ϕ by adding a suitable penalty function to (3.4). For example, LASSO (Tibshirani, 1996) produces a sparse estimator of ϕ via L_1-penalization,

$$\hat{\phi}_{\text{LASSO}} = \arg\min_{\phi} \left\{ -l_n(\phi) + \lambda \sum_{j=1}^{p} |\phi_i| \right\}, \tag{3.5}$$

where $\lambda \geq 0$ is a deterministic parameter to control the degree of penalization. Thus, we can easily obtain a penalized PS estimate of θ by solving (3.3) for given $\hat{\phi}_{\text{LASSO}}$. However, despite the utility of the penalized likelihood method, its applicability is limited to the point estimation in the PS

method. In particular, the derivation of the variance estimator of $\hat{\theta}_{\mathrm{PS}}$ is very challenging under the penalization approach. All the aforementioned concerns motivate us to tackle the sparse propensity estimation problem in a Bayesian framework. We propose to incorporate Bayesian stochastic variable search and approximate Bayesian computation (Beaumont et al., 2002; Soubeyrand and Haon-Lasportes, 2015) into the sparse propensity score estimation. The details are discussed in the following section.

3.3 Bayesian Sparse Propensity Score Estimation

To formulate our proposal, we first introduce a latent variable $z = (z_1, z_2, \cdots, z_p)^{\mathrm{T}}$, which indicates nonzero elements of ϕ as follows:

$$z_j = \begin{cases} 1 & \text{if } \phi_j \neq 0 \\ 0 & \text{if } \phi_j = 0 \end{cases}, \ j = 1, 2, \ldots, p. \tag{3.6}$$

Thus, z_j is an indicator function for the inclusion of j-th covariate into the PS model. Then, by assigning suitable prior distributions for the parameter ϕ and the latent variable z, we can obtain the marginal posterior distribution $p(z|x, \delta)$ using the likelihood of ϕ in (3.4), where $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $\delta = (\delta_1, \delta_2, \ldots, \delta_n)^{\mathrm{T}}$. After the posterior distribution $p(z \mid x, \delta)$ is obtained, we can employ the Bayesian method of Sang and Kim (2017) to generate the posterior distribution of θ, given the response model.

To account for the sparsity of the response model, we assign the Spike-and-Slab Gaussian mixture prior for ϕ and independent Bernoulli prior for z as follows:

$$\phi_j | z_j \overset{ind}{\sim} \mathrm{N}(0, \nu_0(1 - z_j) + \nu_1 z_j), \tag{3.7}$$

$$z_j \overset{ind}{\sim} \mathrm{Ber}(w_j), \tag{3.8}$$

where $w_j (\in (0, 1))$, $\nu_0 (> 0)$, and $\nu_1 (> \nu_0)$ are deterministic hyperparameters. To induce sparsity for ϕ, the scale hyperparameters ν_0 and ν_1 need to be small and large fixed values, respectively. In our simulation study, we set $\nu_0 = 10^{-7}$ and $\nu_1 = 10^{7}$ for $n = 500$. The mixing probability w_j can be interpreted as the prior probability that ϕ_j is nonzero. Under the absence of prior information for ϕ, we can set $w_j = 0.5$ for all j or set the uniform prior for w_j.

Let $L_1(\phi|x,\delta)$ be the likelihood of ϕ obtained from (3.4). Then, our proposed Bayesian sparse propensity score (BSPS) method can be described as following two steps:

Step 1: Generate z^* from the marginal posterior distribution of z given x and δ:

$$z^* \sim p(z|x,\delta) = \frac{\int L_1(\phi|x,\delta)p(\phi|z)p(z)d\phi}{\int\int L_1(\phi|x,\delta)p(\phi|z)p(z)d\phi dz}, \tag{3.9}$$

where $p(\phi|z)$ and $p(z)$ are the prior density functions of ϕ and z, respectively, as defined in (3.7) and (3.8).

Step 2: Generate θ^* from an approximate posterior distribution of θ given the z^* generated from **Step 1**.

We first discuss **Step 1**. To generate z^* from (3.9) in **Step 1** efficiently, the data augmentation algorithm (Wei and Tanner, 1990) can be applied. That is, the marginal posterior distribution of z given x and δ can be obtained by iterating the following two steps until convergence:

I-step: Given ϕ^*, generate z^* from

$$\begin{aligned} z^* \sim p(z|x,\delta,\phi^*) &= \frac{L_1(\phi^*|x,\delta)p(\phi^*|z)p(z)}{\int L_1(\phi^*|x,\delta)p(\phi^*|z)p(z)dz} \\ &= \frac{p(\phi^*|z)p(z)}{\int p(\phi^*|z)p(z)dz} = p(z|\phi^*). \end{aligned}$$

P-step: Given z^*, generate ϕ^* from

$$\phi^* \sim p(\phi|x,\delta,z^*) = \frac{L_1(\phi|x,\delta)p(\phi|z^*)}{\int L_1(\phi|x,\delta)p(\phi|z^*)d\phi}. \tag{3.10}$$

Note that **I-step** and **P-step** perform the model sampling and the parameter sampling, respectively. Under (3.7) and (3.8), **I-step** can be simplified as generating $z^* = (z_1^*, z_2^*, \ldots, z_p^*)^{\mathrm{T}}$ from

$$z_j^* \overset{ind}{\sim} \mathrm{Ber}\left(\frac{w_j\psi(\phi_j^*|0,\nu_1)}{w_j\psi(\phi_j^*|0,\nu_1) + (1-w_j)\psi(\phi_j^*|0,\nu_0)}\right),\ j = 1,2,\ldots,p,$$

where $\psi(\cdot|\mu,\sigma^2)$ denotes a Gaussian density function with mean μ and variance σ^2. Thus, **I-Step** can be efficiently generated.

Note that the normalizing constant in **P-Step** (3.10) is not tractable. To generate ϕ^* from $p(\phi|x, \delta, z^*)$, the Metropolis-Hastings algorithm (Chib and Greenberg, 1995) can be applied. However, when the dimensionality of ϕ is high, the computation can be quite heavy. Thus, instead of using the likelihood function from (3.4), we propose to use the approximate Bayesian computation (ABC) method by treating the estimating equations as the summary statistics for ϕ and using its sampling distribution to replace the likelihood function. The details are given in **Remark 1**.

Remark 3.1. *To discuss the proposed ABC method for approximating (3.10), we define:*

$$S_n(\phi) \quad = \quad n^{-1} \sum_{i=1}^{n} S(\phi; x_i, \delta_i), \tag{3.11}$$

where $S(\phi; x_i, \delta_i) = \partial \log f(\delta_i|x_i, \phi)/\partial\phi$. *Let* $\hat{\phi} = \hat{\phi}(x, \delta)$ *be the solution to the estimating equation* $S_n(\phi) = 0$. *Then, under some regularity conditions, we can establish the asymptotic distribution of* $\hat{\phi}$*:*

$$\sqrt{n}\left(\hat{\phi} - \phi\right) \Big| \, \phi \xrightarrow{d} N_p(0, \Sigma), \tag{3.12}$$

as $n \to \infty$, *where "*$\xrightarrow{d}$*" represents "convergence in distribution" and* Σ *is the covariance matrix of* $\sqrt{n}\hat{\phi}$. *From (3.12), we can get*

$$\hat{\phi}|\phi \sim N_p\left(\phi, n^{-1}\hat{\Sigma}\right), \tag{3.13}$$

where $\hat{\Sigma}$ *is a consistent variance estimator of* Σ. *See Appendix B for the derivation of* $\hat{\Sigma}$. *Let* $g(\hat{\phi}|\phi)$ *be the sampling density function of* $\hat{\phi}$ *in (3.13). For sufficiently large* n, $L_1(\phi|x, \delta)$ *is (approximately) proportional to* $g(\hat{\phi}|\phi)$ *with respect to* ϕ. *Thus, the posterior distribution in (3.10) can be approximated by*

$$p_g(\phi \mid x, \delta, z^*) = \frac{g(\hat{\phi} \mid \phi)p(\phi \mid z^*)}{\int g(\hat{\phi} \mid \phi)p(\phi \mid z^*)d\phi}. \tag{3.14}$$

Since our prior distribution of ϕ *is conjugate for Gaussian distribution, our new algorithm for* ***P-step*** *can be explicitly expressed as follows:*

New P-step: *given* z^*, *generate* ϕ^* *from*

$$\phi^* \mid z^* \sim N_p\left\{\left(\hat{\Sigma}^{-1} + n^{-1}V_{z^*}^{-1}\right)^{-1}\hat{\Sigma}^{-1}\hat{\phi}, \left(n\hat{\Sigma}^{-1} + V_{z^*}^{-1}\right)^{-1}\right\}, \tag{3.15}$$

where $V_{z^*} = Diag\left(\nu_{z_1^*}, \nu_{z_2^*}, \ldots, \nu_{z_p^*}\right)$ *and* $\nu_{z_j^*} = \nu_1 z_j^* + \nu_0(1 - z_j^*)$.

Note that, in **New P-step** (3.15), only V_{z^*} involves z^*. Thus, the **New P-step** does not involves Markov Chain Monte Carlo (MCMC) method and the computation can be very efficient. Furthermore, the **New P-step** is reversible of z^*, in the sense that even some z_j^* occurrently are 0, the proposed method can make z_j^* return to 1, if the true z_j^* are 1. Thus, the proposed method is invariant with the starting point of z^*.

We now discuss **Step 2**. In **Step 2**, to generate the posterior distribution of θ given z^*, we can apply the method of Sang and Kim (2017). Let x_{i,z^*} be a sub-vector of x_i corresponding to the nonzero elements of z^* for $i = 1, 2, \ldots, n$. Similarly, let ϕ_{z^*} be a sub-vector of ϕ corresponding to the nonzero elements of z^*. Given z^*, the joint estimating equations are

$$U_n(\theta, \phi_{z^*}) = \begin{pmatrix} n^{-1}\sum_{i=1}^n S(\phi_{z^*}; x_{i,z^*}, \delta_i) \\ n^{-1}\sum_{i=1}^n \delta_i \pi^{-1}(x_{i,z^*}; \phi_{z^*}) U(\theta; x_i, y_i) \end{pmatrix}, \tag{3.16}$$

where $S(\phi_{z^*}; x_{i,z^*}, \delta_i) = \partial \log f(\delta_i | x_{i,z^*}, \phi_{z^*})/\partial \phi_{z^*}$. Let $\hat{\phi}_{z^*} = \hat{\phi}(x, \delta, z^*)$ and $\hat{\theta}_{z^*} = \hat{\theta}(x, y_{\text{obs}}, \delta, z^*)$ be the solutions to the joint estimating equation $U_n(\theta, \phi_{z^*}) = \mathbf{0}$ in (3.16). Then, **Step 2** can be implemented by generating θ^* from

$$\theta^* \sim p(\theta | x, y_{\text{obs}}, \delta, z^*) = \frac{\int g\left\{U_n(\theta, \phi_{z^*}) | \theta, \phi_{z^*}\right\} p(\theta) p(\phi_{z^*}) d\phi_{z^*}}{\int\int g\left\{U_n(\theta, \phi_{z^*}) | \theta, \phi_{z^*}\right\} p(\theta) p(\phi_{z^*}) d\phi_{z^*} d\theta}, \tag{3.17}$$

where $g\left\{U_n(\theta, \phi_{z^*}) | \theta, \phi_{z^*}\right\}$ is the asymptotic distribution of the joint estimating equations (3.16) and $p(\theta) \times p(\phi_{z^*})$ is the prior distribution for the parameters (θ, ϕ_{z^*}).

The algorithm for generating θ^* from (3.17) without using Taylor linearizion can be implemented in the following two-step procedure.

1. Generate $\eta^* = (\eta_1^*, \eta_2^*)$ from

$$\eta^* \sim \mathrm{N}_{|z^*|+1}\left(0, n^{-1}\hat{\Sigma}_{z^*}\right),$$

where $|z^*| = \sum_{j=1}^p z_j^*$ and

$$\hat{\Sigma}_{z^*} = \frac{1}{n}\left[\begin{array}{c|c} \sum_{i=1}^n S(\hat{\phi}_{z^*}; x_{i,z^*}, \delta_i)^{\otimes 2} & \sum_{i=1}^n \delta_i \hat{\pi}_{i,z^*}^{-1} U(\hat{\theta}_{z^*}; x_i, y_i) S(\hat{\phi}_{z^*}; x_{i,z^*}, \delta_i) \\ \hline \text{symm.} & \sum_{i=1}^n \delta_i \hat{\pi}_{i,z^*}^{-2} \{U(\hat{\theta}_{z^*}; x_i, y_i)\}^2 \end{array}\right],$$

where $\hat{\pi}_{i,z^*} = \pi(x_{i,z^*}; \hat{\phi}_{z^*})$ and $A^{\otimes 2} = AA^{\mathrm{T}}$ for a generic matrix A.

2. Obtain $(\phi^*_{z^*}, \theta^*)$ by solving $U_n(\theta, \phi_{z^*}) = \eta^*$.

Note that, since $g(U_n \mid \theta, \phi_{z^*})$ is a normal distribution, the proposed algorithm also does not involve the MCMC method and the computation is very fast. Sang and Kim (2017) gave a rigorously theoretical justification of the Bayesian method in **Step 2** when the propensity model is correctly specified.

Let $\{\theta^*_{(k)} : k = 1, 2, \ldots, M\}$ be the posterior sample of size M generated from the method we have proposed. Then, our Bayesian sparse propensity score (BSPS) estimator of θ is obtained by

$$\hat{\theta}_{\text{BSPS}} = \sum_{k=1}^{M} \theta^*_{(k)}/M.$$

The $\alpha/2$ and $(1-\alpha/2)$ quantiles of $\{\theta^*_{(k)} : k = 1, 2, \ldots, M\}$ can be directly used to construct a $(1-\alpha)$ credible interval for θ.

3.4 Asymptotic Properties

To establish the asymptotic properties of our BSPS method, we first show the existence of the unique solution to the estimating equation in (3.11). Silvapulle (1981) established the necessary and sufficient conditions for the existence and uniqueness of MLE for binary response models. Let F_1 and F_0 be the relative interiors of the convex cones generated by $x_1, x_2, \cdots, x_n$, that is,

$$F_1 = \left\{ \sum_{i=1}^{n} \delta_i k_i x_i : k_i > 0 \right\} \text{ and } F_0 = \left\{ \sum_{i=1}^{n} (1-\delta_i) k_i x_i : k_i > 0 \right\}.$$

Similar to Silvapulle (1981), we consider the following underlying assumptions.

(**A1**) Let $X_n = [x_1, \cdots, x_n]^{\mathrm{T}}$. Assume $X_n^{\mathrm{T}} X_n$ is a full rank design matrix.

(**A2**) Let x_{i1} be the first element of x_i for $i = 1, 2, \ldots, n$. Assume $x_{i1} = 1$ for all i.

(**A3**) Suppose that $-\log G$ and $\log(1-G)$ are convex. Further assume that G is strictly increasing for t such that $0 < G(t) < 1$.

Assumption (**A1**) requires the design matrix to be full rank, which is a common assumption in the linear regression setup. If not, we can remove redundant variables to make (**A1**) satisfied.

Assumption (**A2**) means the intercept is included in the model. The most common link functions, such as logit and probit functions, satisfy assumption (**A3**).

Lemma 3.2. *Under (**A1**) – (**A3**) and the MAR assumption in (3.2), the solution to the estimating equation in (3.11) is uniquely defined if and only if* $F_1 \cap F_0 \neq \emptyset$*, where* $\emptyset$ *represents the empty set.*

The proof of Lemma 3.2 can be found in Silvapulle (1981). In context of the PS estimation, condition $F_1 \cap F_2 \neq \emptyset$ is satisfied if the response probability is bounded below as in Rosenbaum (1987) and Kim and Kim (2007). Thus, Lemma 3.2 is also necessary in the PS estimation. Once the MLE of ϕ exists, the asymptotic distribution $g(\hat{\phi} \mid \phi)$ can be used to approximate the likelihood function $L_1(\phi \mid x, \delta)$ and the posterior distribution of z can be derived in a closed form.

We now establish the model selection consistency under the Bayesian framework. The Bayesian model selection consistency is satisfied if the posterior probability of the true model tends to one as the sample size goes to infinity (Casella et al., 2009). To achieve the model selection consistency or Oracle property (Fan and Li, 2001; Zou, 2006) , we further assume the following conditions.

(**A4**) Assume $p = O(1)$, where p is the dimension of ϕ (or X).

(**A5**) For the hyperparameters, assume that $\nu_0 = o(n^{-1})$, $\nu_1 = O(1)$, and $w_1 = w_2 = \cdots = w_p = 0.5$.

(**A6**) The $\hat{\Sigma}$ in (3.13) satisfies $\hat{\Sigma} = \Sigma\{1 + o_p(1)\}$.

Note that, we assume that p is large but does not dependent on n in assumption (**A4**). Since the approximated sampling distribution $g(\hat{\phi} \mid \phi)$ has the variance of $O_p(n^{-1})$, $\nu_0 = o(n^{-1})$ and $\nu_1 = O(1)$ are in the right scales to approximate the Spike-and-Slab prior in assumption (**A5**). The choice of $w_j = 0.5$ represents a non-informative prior for each covariate component. Assumption (**A6**) requires that the variance covariance estimator be consistent to make the approximation of the sampling distribution valid. The following theorem establishes the oracle property of the proposed BSPS method.

Theorem 3.1. *Under assumptions (**A1**)–(**A6**) and the MAR assumption in (3.2), we have*

$$p_g(z_o|x,\delta) \to 1,$$

in probability, where z_o is the true model and

$$p_g(z|x,\delta) = \frac{\int g(\hat{\phi}|\phi)p(\phi|z)p(z)d\phi}{\int\int g(\hat{\phi}|\phi)p(\phi|z)p(z)d\phi dz}.$$

The proof of Theorem 3.1 is given in Appendix C. According to Theorem 3.1, we observe that the probability that **Step 1** selects the true model becomes very close to one when the sample size n is sufficiently large. Thus, the proposed Bayesian method can effectively eliminate irrelevant covariates and select important ones to adjust bias due to nonresponse. Since we assume the true response model is sparse, $p_o = \sum_j z_{o,j}$ is relatively small compared to n. Thus, the asymptotic normality is easy to establish under the regularity conditions.

Corollary 3.1. *Under the conditions in Theorem 3.1 and the regularity conditions of Sang and Kim (2017), we have*

$$\left\{\hat{V}ar(\hat{\theta}_{BSPS})\right\}^{-1/2}\left(\hat{\theta}_{BSPS}-\theta_0\right) \xrightarrow{d} N(0,1),$$

where θ_0 satisfies $E\{U(\theta;X,Y)\}=0$ and

$$\hat{V}ar(\hat{\theta}_{BSPS}) = \sum_{k=1}^{M}\left(\theta^*_{(k)}-\hat{\theta}_{BSPS}\right)^2/(M-1).$$

Sang and Kim (2017) have established the asymptotic normality of the Bayesian propensity score (BPS) estimator under the correctly specified response model. By Theorem 3.1, the probability that **Step 1** selects the true model converges to one. Consequently, the asymptotic distribution of our BSPS estimator is the same as the asymptotic distribution of BPS estimator under the true model which leads to the asymptotic normality of the BSPS estimator.

Remark 3.2. *From Theorem 3.1, we can see that the model uncertainty of z vanishes as $n \to \infty$. However, in the finite sample, the model uncertainty always contributes to the variability of $\hat{\theta}_{BSPS}$. The advantage of the proposed Bayesian method is that it can still capture the variability of the*

model uncertainty in the finite sample case. Since for each $z^* \sim p(z \mid x, \delta)$*, we apply one step algorithm in* ***Step 2****. Thus, by Law of Large Numbers (LLN), we can show that*

$$\begin{aligned}\sum_{k=1}^{M}\left(\theta^*_{(k)} - \hat{\theta}_{BSPS}\right)^2/(M-1) \quad &\xrightarrow{P} Var\left\{\theta^*\right\} \\ &= Var\left\{E\left(\theta^* \mid z^*\right)\right\} + E\left\{Var\left(\theta^* \mid z^*\right)\right\},\end{aligned}$$

where θ^* *is generated from* ***Step 2****. In the finite sample,* $Var\left\{E\left(\theta^* \mid z^*\right)\right\}$ *represents the variability due to the model uncertainty. When* $n \to \infty$*,* $Pr(z^* = z_o) = 1$*, which leads to* $Var\left\{E\left(\theta^* \mid z^*\right)\right\} = 0$*.*

3.5 Simulation Study

In this section, we conduct two simulation studies to examine the finite sample performance of the proposed Bayesian method. The first simulation study investigates the proposed Bayesian method under the IID setup. In the second simulation study, we apply our proposed method using a real data obtained from a probability sampling.

3.5.1 Simulation study I

In the first simulation, our data generation process consists of the following two parts.

1. Generate a random sample of size $n = 500$, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$, from each one of the following two models:

$$\mathcal{M}_1 \quad : \quad y_i \overset{ind}{\sim} 2x_{i1} + 2x_{i2} + e_i; \tag{3.18}$$

$$\mathcal{M}_2 \quad : \quad y_i \overset{ind}{\sim} \text{Binomial}\left\{20, p(x_i)\right\}; \tag{3.19}$$

   where $p(x_i) = \exp(x_{i3})/\left\{1 + \exp(x_{i3})\right\}$, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ with $x_{i1} = 1$, $x_{i2}, x_{i3}, \ldots, x_{ip} \overset{iid}{\sim}$ $\mathrm{N}(0, 1)$, and the errors e_i are generated independently from χ^2_3.

2. For $i = 1, 2, \ldots, n$, generate the response indicator of y_i from each one of the following two response mechanisms:

$$\mathcal{R}_1 \quad : \quad \delta_i \overset{ind}{\sim} \text{Ber}\left\{\frac{\exp(x_{i1} + x_{i2})}{1 + \exp(x_{i1} + x_{i2})}\right\}; \tag{3.20}$$

$$\mathcal{R}_2 \quad : \quad \delta_i \overset{ind}{\sim} \text{Ber}\left\{\Phi(0.7x_{i1} + 0.7x_{i2})\right\}; \tag{3.21}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of $\mathrm{N}(0,1)$. Here, the average response rate is about 0.7 for the both response mechanisms.

We consider all possible cases: (Case 1) $\mathcal{M}_1$ and $\mathcal{R}_1$; (Case 2) $\mathcal{M}_1$ and $\mathcal{R}_2$; (Case 3) $\mathcal{M}_2$ and $\mathcal{R}_1$; (Case 4) $\mathcal{M}_2$ and $\mathcal{R}_2$. In each case, we perform $2,000$ Monte Carlo replications for each $p = 5, 10, 50$ and 100. Note that in our setup p controls the amount of sparsity on the propensity score. As p increases, the propensity score becomes more sparse. We are interested in estimating $\theta = E(Y)$, which is the solution of $E\{U(\theta; X, Y)\} = E(Y - \theta) = 0$. We use a working PS model $G(t) = \exp(t)/\{1 + \exp(t)\}$, which is the true link function in $\mathcal{R}_1$.

For each setup, we generate 500 Monte Carlo samples and for each realized sample, we apply following methods:

1. PS: The traditional PS estimate, say $(\hat{\phi}_{\mathrm{PS}}, \hat{\theta}_{\mathrm{PS}})$, is obtained by solving the joint estimating equations

$$\frac{1}{n}\sum_{i=1}^{n}\{\delta_i - \pi(x_i;\phi)\}\, x_i = 0,$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(x_i;\phi)}(y_i - \theta) = 0,$$

   where $\pi(x_i;\phi) = G(x_i^T\phi)$. The variance of $(\hat{\phi}_{\mathrm{PS}}, \hat{\theta}_{\mathrm{PS}})$ is estimated by the Taylor linearization. The 95% confidence intervals are constructed from the asymptotic normal distribution of $(\hat{\phi}_{\mathrm{PS}}, \hat{\theta}_{\mathrm{PS}})$.

2. TPS: The true propensity score (TPS) method in which the ordinary PS method is applied using the covariates in the true response mechanism. The 95% confidence intervals are constructed from the asymptotic normal distribution of $(\hat{\phi}_{\mathrm{TPS}}, \hat{\theta}_{\mathrm{TPS}})$

3. LASSO: We first apply the LASSO method to select the response model with λ in (3.5) chosen by the cross-validation method. The algorithm is implemented in "glmnet" (Friedman et al., 2009). Then we apply the traditional PS estimation method to the selected response model. Variances and confidence intervals are obtained by using the asymptotic normal distribution of $(\hat{\phi}_{\mathrm{LASSO}}, \hat{\theta}_{\mathrm{LASSO}})$ for the selected response model.

4. BSPS: In BSPS, we set $w_1 = \cdots = w_p = 0.5$, $\nu_0 = 10^{-7}$, and $\nu_1 = 10^7$ to induce noninformative priors. Using the formula in Section 3.3, we compute the BSPS estimate and its variance estimate based on the posterior sample of size $2,000$ after $2,000$ burn-in iterations. The 95% confidence intervals are constructed from the quantiles of the posterior sample.

To assess the variable selection performance of BSPS and LASSO methods, we compute true positive rate (TPR) and true negative rate (TNR), where TPR is the proportion of the regression coefficients that are correctly identified as nonzero and TNR is the proportion of the regression coefficients that are correctly identified as zero. The coverage probabilities of each methods are computed by counting how often the confidence intervals contains the true parameter values. The simulation results for each choice of $(\mathcal{M}, \mathcal{R})$ are presented in Tables 3.1, 3.2, 3.3, and 3.4, respectively.

Table 3.1: Table: Simulation results for Case 1 ($\mathcal{M}_1$,$\mathcal{R}_1$): "Bias" is the bias of the point estimator for θ, "S.E." represents the standard error of the point estimator, "E[S.E.]" is the estimated standard error, "CP" represents the coverage probability of the 95% confidence interval estimate.

p	Method	Bias	S.E.	E[S.E.]	CP	TPR	TNR
5	PS	0.001	0.173	0.168	0.953		
5	TPS	0.001	0.171	0.168	0.952		
5	LASSO	0.001	0.172	0.168	0.952	1.000	0.639
5	BSPS	-0.006	0.173	0.168	0.949	1.000	0.995
10	PS	0.004	0.173	0.168	0.951		
10	TPS	0.003	0.171	0.168	0.951		
10	LASSO	0.003	0.172	0.169	0.952	1.000	0.749
10	BSPS	-0.004	0.172	0.168	0.946	1.000	0.994
50	PS	0.012	0.189	0.161	0.923		
50	TPS	0.004	0.171	0.168	0.955		
50	LASSO	0.007	0.175	0.169	0.956	1.000	0.904
50	BSPS	-0.003	0.173	0.168	0.953	1.000	0.995
100	PS	0.023	0.235	0.147	0.828		
100	TPS	0.007	0.172	0.167	0.947		
100	LASSO	0.012	0.183	0.170	0.944	1.000	0.937
100	BSPS	0.002	0.174	0.168	0.944	0.998	0.996

Table 3.2: Simulation results for Case 2 ($\mathcal{M}_1$,$\mathcal{R}_2$): "Bias" is the bias of the point estimator for θ, "S.E." represents the standard error of the point estimator, "E[S.E.]" is the estimated standard error, "CP" represents the coverage probability of the 95% confidence interval estimate

.

p	Method	Bias	S.E.	E[S.E.]	CP	TPR	TNR
5	PS	0.001	0.166	0.168	0.949		
5	TPS	0.001	0.167	0.169	0.949		
5	LASSO	0.002	0.167	0.169	0.949	1.000	0.627
5	BSPS	-0.005	0.169	0.169	0.949	1.000	0.992
10	PS	0.006	0.176	0.168	0.942		
10	TPS	0.004	0.169	0.169	0.944		
10	LASSO	0.004	0.173	0.169	0.946	1.000	0.737
10	BSPS	-0.004	0.170	0.169	0.944	1.000	0.993
50	PS	0.016	0.190	0.160	0.911		
50	TPS	0.002	0.171	0.168	0.948		
50	LASSO	0.009	0.179	0.170	0.945	1.000	0.897
50	BSPS	-0.006	0.175	0.169	0.948	1.000	0.995
100	PS	0.045	0.223	0.144	0.794		
100	TPS	-0.004	0.176	0.169	0.948		
100	LASSO	0.010	0.181	0.170	0.939	1.000	0.935
100	BSPS	-0.009	0.180	0.169	0.947	0.999	0.995

From Table 3.1, where we correctly specify the link function for the response model, we observe that when p is small $(5, 10)$, the PS, LASSO and BSPS methods work well and provide very similar results to the TPS method. However, in term of the probability of correctly identifying the true response model, the BSPS method always performs better than the LASSO method. As p increases $(50, 100)$, the bias and the variance of PS estimator increase. But, the proposed BSPS method is still consistent and the variance of BSPS estimator does not change with p as in the TPS method. As a result, the coverage probability of the confidence intervals for the PS method is quite poor. Comparing the true standard errors with the estimated standard errors, the PS method and LASSO method are always under-estimate for large p, which confirms that the asymptotic normality of the PS method fails for large p and the LASSO method fails to account for the model uncertainty. Simulation results in Table 3.1 clearly shows that the BSPS method is consistently efficient regardless of changes in p. Note that BSPS successfully eliminates all the

Table 3.3: Simulation results for Case 3 ($\mathcal{M}_2$,$\mathcal{R}_1$): "Bias" is the bias of the point estimator for θ, "S.E." represents the standard error of the point estimator, "E[S.E.]" is the estimated standard error, "CP" represents the coverage probability of the 95% confidence interval estimate.

p	Method	Bias	S.E.	E[S.E.]	CP	TPR	TNR
5	PS	0.010	0.223	0.223	0.954		
5	TPS	0.006	0.258	0.260	0.955		
5	LASSO	0.007	0.230	0.252	0.968	1.000	0.653
5	BSPS	0.005	0.254	0.259	0.958	1.000	0.990
10	PS	-0.006	0.227	0.223	0.946		
10	TPS	-0.003	0.264	0.260	0.952		
10	LASSO	-0.007	0.239	0.255	0.961	1.000	0.749
10	BSPS	-0.004	0.263	0.260	0.953	1.000	0.994
50	PS	0.009	0.249	0.213	0.914		
50	TPS	0.008	0.268	0.260	0.945		
50	LASSO	0.010	0.261	0.261	0.951	1.000	0.904
50	BSPS	0.008	0.267	0.260	0.946	1.000	0.995
100	PS	-0.004	0.285	0.194	0.834		
100	TPS	-0.005	0.264	0.260	0.949		
100	LASSO	0.000	0.262	0.262	0.956	1.000	0.937
100	BSPS	-0.003	0.264	0.260	0.948	0.998	0.996

irrelevant covariates. As a result, the performance of the BSPS method is always comparable to the performance of the TPS method. Table 3.2 shows the simulation result when the parametric model of response mechanism is misspecified. The result shows that our proposed method is still stable and accurate, but the PS performs poorly in large values of p. Even though the LASSO method has around 95% coverage probabilities, the estimated standard errors are under-estimated for large $p = (50, 100)$. From Table 3.3, we observe that our proposed BSPS method works very well even under discrete response variables. Also, we can see that the LASSO method cannot provided consistent estimates for the standard errors and correct confidence intervals, when p is small. Table 3.4 shows the most challenging case in which the parametric model for the response mechanism is misspecified and the outcome regression model is not linear. Nevertheless, our BSPS method is still consistent and comparable to the TPS method and the LASSO method fails to provide accurate estimates of standard errors and confidence intervals, when p is small.

Table 3.4: Simulation results for Case 4 ($\mathcal{M}_2$,$\mathcal{R}_2$): "Bias" is the bias of the point estimator for θ, "S.E." represents the standard error of the point estimator, "E[S.E.]" is the estimated standard error, "CP" represents the coverage probability of the 95% confidence interval estimate.

p	Method	Bias	S.E.	E[S.E.]	CP	TPR	TNR
5	PS	-0.002	0.228	0.225	0.949		
5	TPS	-0.002	0.261	0.260	0.951		
5	LASSO	-0.007	0.235	0.251	0.960	1.000	0.628
5	BSPS	-0.002	0.261	0.260	0.949	1.000	0.992
10	PS	0.008	0.229	0.224	0.947		
10	TPS	0.008	0.260	0.259	0.949		
10	LASSO	0.007	0.240	0.254	0.960	1.000	0.743
10	BSPS	0.010	0.259	0.258	0.949	1.000	0.993
50	PS	-0.010	0.247	0.213	0.916		
50	TPS	-0.000	0.266	0.260	0.945		
50	LASSO	0.003	0.258	0.260	0.950	1.000	0.899
50	BSPS	-0.002	0.266	0.260	0.948	1.000	0.995
100	PS	-0.001	0.292	0.191	0.824		
100	TPS	0.005	0.259	0.259	0.950		
100	LASSO	-0.002	0.256	0.261	0.955	1.000	0.935
100	BSPS	0.004	0.259	0.259	0.946	0.999	0.995

3.5.2 Simulation study II

We also apply the proposed Bayesian method to the 2006 Korean Labor and Income Panel Survey (KLIPS) data. A breif description of the panel survey can be found at http://www.kli.re.kr/klips/en/about/introduce.jsp. In KLIPS data, there are 2,506 regular wage earners. The study variable y is the monthly income in 2006. The auxiliary variables (x) include the average monthly income in previous year and demographic variables. The auxiliary variable x is briefly described in Table 3.5.

We grouped age into three groups: age $< 35, 35 \leq$ age $< 51,$ age ≥ 51. We also standardized the continuous auxiliary variable by subtracting its mean and dividing its standard error. Note that the dimension of x is not so large. To demonstrate the proposed Bayesian sparse propensity method, we add additional 50 auxiliary variables as noise variables. Thus, $x = (x_1, \cdots, x_9, x_{10}, \cdots x_{59})^T$, where $(x_1, \cdots, x_9)$ are the auxiliary variables in Table 3.5 and $(x_{10}, \cdots, x_{59})^T \sim N(0, I_p)$ where

Table 3.5: Levels of each auxiliary variable.

variable	levels
gender (x_1)	2
age (x_2)	3
level of education (x_3)	8
job type (x_4)	2
occupation (x_5)	10
maternity leave (x_6)	3
private pension (x_7)	3
labor union (x_8)	3
average monthly income in the previous year (Korean Won 10,000) (x_9)	continuous

$p = 50$ and I_p is a p-dimensional identity matrix. In this simulation study, we use the KLIPS data as a finite population. The realized sample is then obtained from the population by Simple Random Sampling (SRS) with sample size $n = 500$ independently. Since the KLIPS data are fully observed data, we artificially create nonresponse data by applying some missing mechanism. Note that, there are two major differences with the first simulation study. One is the mixed data types of the auxiliary variables. Another is that the outcome regression model is unknown. The simulation process is described as following:

Step 1: Obtain 500 samples from the KLIPS data by SRS.

Step 2: Apply the response mechanism $\mathcal{R}$ to the sample, where the auxiliary variables are fully observed and the study variable y is subject to missingness.

Step 3: Apply the PS method and the proposed Bayesian method to the incomplete sample.

Step 4: Repeat Step 1–3 for $B = 2,000$ times.

The true response function $\mathcal{R}$ is

$$Pr(\delta_i = 1 \mid \boldsymbol{x}_i, y_i) = \frac{\exp(\phi_0 + \phi_1 I_{\{x_{i1}=1\}} + \phi_2 x_{i9})}{1 + \exp(\phi_0 + \phi_1 I_{\{x_{i1}=1\}} + \phi_2 x_{i9})},$$

where $(\phi_0, \phi_1, \phi_2) = (0, 1, 1)$, $I_{\{\cdot\}}$ is an indicator function and the response rate is approximately 65%. Suppose we are interested in the average monthly income $\theta = E(y)$. Therefore, the estimating

equation is $U(\theta; x, y) = y - \theta$. Also, we are interested in Gini coefficient G. The Gini coefficient is an important index of the income inequality, which is also known as Gini index or Gini ratio. The Gini coefficient measures the income distribution and inequality. The Gini coefficient can be calculated by solving

$$\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j| - 2nG\sum_{i=1}^{n} y_i = 0,$$

if y_i are fully observed.

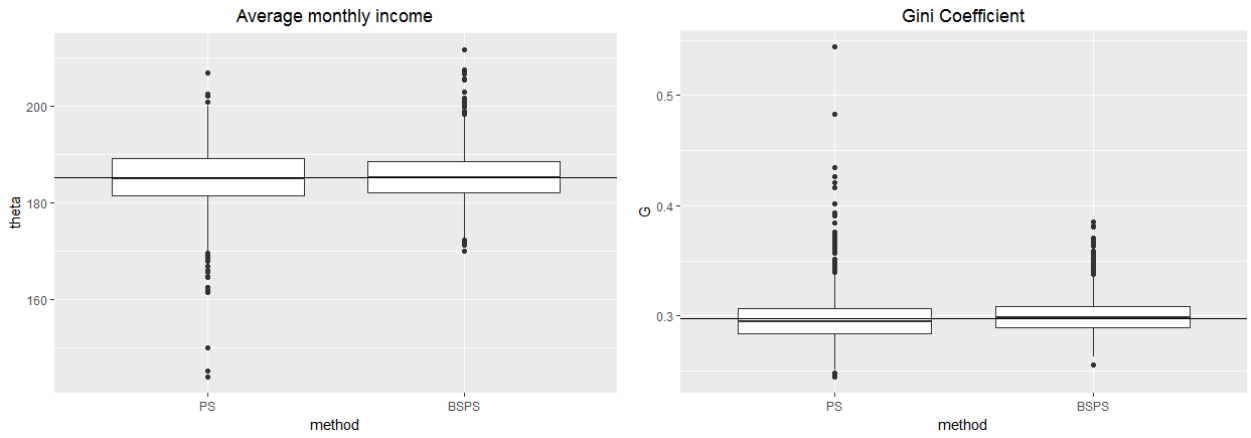
To fit the response model, we assume the response mechanism is

$$Pr(\delta_i = 1 \mid x_i, y_i) = \frac{\exp(x_i^T\phi)}{1+\exp(x_i^T\phi)} =: \pi(\phi; x_i),$$

which is known up to the parameter ϕ. Thus, the joint estimating equations are

$$U_n(\phi, \theta, G) = \begin{cases} n^{-1}\sum_{i=1}^{n}\{\delta_i - \pi(\phi; x_i)\}\, x_i \\ n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\phi;x_i)}(y_i - \theta) \\ n^{-2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j}{\pi(\phi;x_i)\pi(\phi;x_j)}|y_i - y_j| - 2nG\sum_{i=1}^{n}\frac{\delta_i}{\pi(\phi;x_i)}y_i\right\}. \end{cases} \tag{3.22}$$

We apply the PS method and the proposed Bayesian method to (3.22). The analysis result is summarized in Figure 3.1.



(a) Estimated average monthly incomes. The horizontal line is the true population mean.

(b) Estimated Gini coefficients. The horizontal line is the true Gini coefficient in the population.

Figure 3.1: Simulation results for the PS and BSPS methods

From Figure 3.1, we can see that both methods are consistent, but the proposed BSPS method is more efficient than the PS method because of accounting for the response model sparsity. In average monthly income estimation, the PS method provides some extremely small estimates. Also, in the Gini coefficient estimation, the PS method has some extremely large estimates. This is because the PS method including all covariates involves computing inversion of high dimensional covariance matrix and the convergence is not guaranteed. Thus, the PS estimator is significantly affected if the some estimated propensity scores are close to 0. Because of accounting for the sparsity, the BSPS method avoids this situation.

3.6 Discussion

Bayesian approach to PS estimation using the Spike-and-Slab prior for the response propensity model is proposed. In the proposed method, model selection consistency holds and the uncertainty in the model selection is fully captured in the Bayesian framework. The approach provides valid frequentist coverage probabilities in finite samples. Since the PS estimation is widely used in causal inference (Morgan and Winship, 2014; Hudgens and Halloran, 2008), applying the proposed method to the sparse Bayesian causal inference can be developed similarly. Also, our proposed method is developed under the assumption of MAR. Extension of our proposed method to nonignorable nonresponse is a topic for future research.

There are three Appendices in supplementary materials. Appendix A is the proof for Lemma 3.1. In Appendix B, we show how to derive the consistenct variance estimator in (3.13). We present the proof of Theorem 3.1 in Appendix C.

3.7 Appendix A: Proof of Lemma 3.1

To formulate the problem, denote

$$U_1(\theta) = \sum_{i=1}^{n} \frac{\delta_i}{Pr(\delta_i = 1 \mid X_{1i}, X_{i2})} U(\theta; X_{1i}, y_i),$$
$$U_2(\theta) = \sum_{i=1}^{n} \frac{\delta_i}{Pr(\delta_i = 1 \mid X_{1i})} U(\theta; X_{1i}, y_i).$$

By Taylor linearization and ignoring the smaller term, we can obtain

$$Var\left(\hat{\theta}_{PS}\right) = \left[E\left\{\frac{\partial U_1(\theta)}{\partial \theta}\right\}\right]^{-1} Var\left\{U_1(\theta)\right\} \left[E\left\{\frac{\partial U_1(\theta)}{\partial \theta}\right\}\right]^{-1}, \tag{3.23}$$
$$Var\left(\hat{\theta}_{SPS}\right) = \left[E\left\{\frac{\partial U_2(\theta)}{\partial \theta}\right\}\right]^{-1} Var\left\{U_2(\theta)\right\} \left[E\left\{\frac{\partial U_2(\theta)}{\partial \theta}\right\}\right]^{-1}. \tag{3.24}$$

See Chapter 5 in Kim and Shao (2013) for details. Note that

$$\begin{aligned} E\left\{\tfrac{\partial U_1(\theta)}{\partial \theta}\right\} &= \sum_{i=1}^{n} E\left[E\left\{\frac{\delta_i}{Pr(\delta_i = 1 \mid X_{1i}, X_{i2})} \frac{\partial U(\theta; X_{1i}, y_i)}{\partial \theta} \middle| X_i, y_i\right\}\right] \\ &= \sum_{i=1}^{n} E\left\{\frac{\partial U(\theta; X_{1i}, y_i)}{\partial \theta}\right\}, \end{aligned}$$

and

$$\begin{aligned} E\left\{\tfrac{\partial U_2(\theta)}{\partial \theta}\right\} &= \sum_{i=1}^{n} E\left[E\left\{\frac{\delta_i}{Pr(\delta_i = 1 \mid X_{1i})} \frac{\partial U(\theta; X_{1i}, y_i)}{\partial \theta} \middle| X_{i1}, y_i\right\}\right] \\ &= \sum_{i=1}^{n} E\left\{\frac{\partial U(\theta; X_{1i}, y_i)}{\partial \theta}\right\}. \end{aligned}$$

Thus, from equations (3.23) and (3.24), to show $Var\left(\hat{\theta}_{PS}\right) \geq Var\left(\hat{\theta}_{SPS}\right)$ is equivalent to showing $Var\left\{U_1(\theta)\right\} \geq Var\left\{U_2(\theta)\right\}$.

Now, we derive the variance of $U_1(\theta)$ and $U_2(\theta)$, respectively. Since we have shown that $Var\{U_1(\theta)\} \geq Var\{U_2(\theta)\}$ implies $Var\left(\hat{\theta}_{PS}\right) \geq Var\left(\hat{\theta}_{SPS}\right)$, it is sufficient to derive follows:

$$\begin{aligned} Var\{U_1(\theta)\} &= Var\left[E\{U_1(\theta) \mid X_i, y_i\}\right] + E\left[Var\{U_1(\theta) \mid X_i, y_i\}\right] \\ &= Var\left\{\sum_{i=1}^{n} U(\theta; X_{i1}, y_i)\right\} \\ &+E\left\{\sum_{i=1}^{n} \frac{1 - Pr(\delta_i = 1 \mid X_{i1}, X_{i2})}{Pr(\delta_i = 1|X_{i1}, X_{2i})} U^2(\theta; X_{i1}, y_i)\right\}, \qquad (3.25) \\ Var\{U_2(\theta)\} &= Var\left[E\{U_2(\theta) \mid X_{i1}, y_i\}\right] + E\left[Var\{U_2(\theta) \mid X_{i1}, y_i\}\right] \\ &= Var\left\{\sum_{i=1}^{n} U(\theta; X_{i1}, y_i)\right\} \\ &+E\left\{\sum_{i=1}^{n} \frac{1 - Pr(\delta_i = 1 \mid X_{i1})}{Pr(\delta_i = 1|X_{i1})} U^2(\theta; X_{i1}, y_i)\right\}, \qquad (3.26) \end{aligned}$$

By Jensen's inequality, we have

$$\begin{aligned} E\left\{ \frac{1}{Pr(\delta_i=1|X_{1i}, X_{2i})} \middle| X_{1i} \right\} &\geq \frac{1}{E\left\{ Pr(\delta_i = 1 \mid X_{1i}, X_{2i}) \middle| X_{1i} \right\}} \\ &= \frac{1}{Pr(\delta_i = 1|X_{1i})}. \qquad (3.27) \end{aligned}$$

Therefore, combining (3.25),(3.26) and (3.27), we have $Var\{U_1(\theta)\} \geq Var\{U_2(\theta)\}$. Thus, $Var\left(\hat{\theta}_{PS}\right) \geq Var\left(\hat{\theta}_{SPS}\right)$ holds, which completes the proof.

3.8 Appendix B: Consistent variance estimator of Σ

Since $\hat{\phi}$ is the solution to

$$S_n(\phi) \quad = \quad n^{-1} \sum_{i=1}^{n} S(\phi; x_i, \delta_i),$$

and according to Theorem 5.21 in Van der Vaart (2000), we have

$$\Sigma/n = A^{-1} B \left(A^{\mathrm{T}}\right)^{-1},$$

where $A = E\left\{\frac{\partial S_n(\phi)}{\partial \phi}\right\}$ and $B = Var\{S_n(\phi)\}$. Hence, using the Law of Large Numbers (LLN), we can obtain a consistent variance estimator of Σ as

$$\hat{\Sigma}/n = \hat{A}^{-1} \hat{B} \left(\hat{A}^{\mathrm{T}}\right)^{-1},$$

where the consistent estimators $\hat{A}$ and $\hat{B}$ can be obtained by

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial S_n(\phi)}{\partial \phi}\bigg|_{\phi=\hat{\phi}} \quad \text{and} \quad \hat{B} = \frac{1}{n}\sum_{i=1}^{n} S_n(\hat{\phi}) S_n(\hat{\phi})^{\mathrm{T}}.$$

3.9 Appendix C: Proof of Theorem 3.1

Let $V = n^{-1}\Sigma$. Our proof can be summarized as follows: First, we show that

$$\tilde{p}(z_o|x,\delta) \xrightarrow{p} 1, \tag{3.28}$$

as $n \to \infty$, where

$$\tilde{p}(z_o|x,\delta) = \frac{\int \psi(\hat{\phi}|\phi, V) p(\phi|z_o) p(z_o) d\phi}{\int\int \psi(\hat{\phi}|\phi, V) p(\phi|z) p(z) d\phi dz},$$

and $\psi(\cdot \mid \phi, V)$ is the normal density function with mean ϕ and variance V. Second, we show that

$$|\tilde{p}(z_o|x,\delta) - p_g(z_o|x,\delta)| \xrightarrow{p} 0, \tag{3.29}$$

as $n \to \infty$. Note that

$$|\tilde{p}(z_o|x,\delta) - p_g(z_o|x,\delta)| \geq \left| |\tilde{p}(z_o|x,\delta) - 1| - |p_g(z_o|x,\delta) - 1| \right|.$$

Finally, by (3.28) and (3.29), we have that

$$p_g(z_o|x,\delta) \xrightarrow{p} 1,$$

as $n \to \infty$.

Proof of Claim (3.28)

Under (**A5**), since $\pi(z) \propto 1$, $\tilde{p}(z_o|x,\delta)$ reduces to

$$\begin{aligned}\tilde{p}(z_o|x,\delta) &= \frac{\int \psi(\hat{\phi}|\phi,V)p(\phi|z_o)d\phi}{\sum_{z\in\{0,1\}^p}\int \psi(\hat{\phi}|\phi,V)p(\phi|z)d\phi} \\ &:= \frac{f(\hat{\phi}|z_o)}{\sum_{z\in\{0,1\}^p} f(\hat{\phi}|z)} \\ &= \frac{1}{1+\sum_{z\neq z_o}\frac{f(\hat{\phi}|z)}{f(\hat{\phi}|z_o)}},\end{aligned}$$

where $f(\hat{\phi}|z) = \int \psi(\hat{\phi}|\phi,V)p(\phi|z)d\phi$. Under (**A4**), our proof can be done by showing that for any $z \neq z_o$,

$$\frac{f(\hat{\phi}|z)}{f(\hat{\phi}|z_o)} \xrightarrow{p} 0, \tag{3.30}$$

as $n \to \infty$. Since Σ is symmetric and positive definite, by the spectral decomposition, Σ can be factorized as $\Sigma = Q\Lambda Q^{-1}$, where Λ is the diagonal matrix whose diagonal elements are the eigenvalues of Σ and each column of Q is the eigenvector of Σ. Since $V = n^{-1}\Sigma$, we have $V = Q(n^{-1}\Lambda)Q^{-1}$. Let $\lambda_{n,\min} = n^{-1}\lambda_{\min}$ and $\lambda_{n,\max} = n^{-1}\lambda_{\max}$, where $\lambda_{\min}$ and $\lambda_{\max}$ indicate the smallest and the largest diagonal elements of Λ, respectively. Note that $\lambda_{n,\min}^{-1}I - V^{-1}$ and $V^{-1} - \lambda_{n,\max}I$ are positive semidefinite due to the fact that

$$\begin{aligned}\lambda_{n,\min}^{-1}I - V^{-1} &= Q\left(\lambda_{n,\min}^{-1}I - n\Lambda^{-1}\right)Q^{-1}, \\ V^{-1} - \lambda_{n,\max}^{-1}I &= Q\left(n\Lambda^{-1} - \lambda_{n,\max}^{-1}I\right)Q^{-1}.\end{aligned}$$

This implies that

$$\lambda_{n,\max}^{-1}w^{\mathrm{T}}w \leq w^{\mathrm{T}}V^{-1}w \leq \lambda_{n,\min}^{-1}w^{\mathrm{T}}w, \tag{3.31}$$

for any w. Recall that

$$\psi(\hat{\phi}|\phi,V) = c\exp\left\{-\frac{1}{2}\left(\hat{\phi}-\phi\right)^{\mathrm{T}}V^{-1}\left(\hat{\phi}-\phi\right)\right\},$$

where c denotes the normalizing constant. From (3.31), we have

$$\psi(\hat{\phi}|\phi, V) \geq c \exp\left\{-\sum_{j=1}^{p} \frac{1}{2\lambda_{n,\min}} \left(\hat{\phi}_j - \phi_j\right)^2\right\}, \tag{3.32}$$

$$\psi(\hat{\phi}|\phi, V) \leq c \exp\left\{-\sum_{j=1}^{p} \frac{1}{2\lambda_{n,\max}} \left(\hat{\phi}_j - \phi_j\right)^2\right\}. \tag{3.33}$$

Using (3.32), we construct a lower bound of $f(\hat{\phi}|z) = \int \psi(\hat{\phi}|\phi, V)p(\phi|z)d\phi$ as

$$\begin{aligned} f(\hat{\phi}|z) &\geq c \prod_{j=1}^{p} \left(2\pi\nu_{z_j}\right)^{-1/2} \int \exp\left\{-\frac{1}{2\lambda_{n,\min}} \left(\hat{\phi}_j - \phi_j\right)^2 - \frac{1}{2\nu_{z_j}}\phi_j^2\right\} d\phi_j \\ &= c_2 \prod_{j=1}^{p} \left(\frac{\lambda_{n,\min}}{\lambda_{n,\min} + \nu_{z_j}}\right)^{1/2} \exp\left\{-\frac{\hat{\phi}_j^2}{2\left(\lambda_{n,\min} + \nu_{z_j}\right)}\right\} \equiv L_f(z). \end{aligned}$$

Similarly, using (3.33), we construct an upper bound of $f(\hat{\phi}|z)$ as

$$f(\hat{\phi}|z) \leq c_3 \prod_{j=1}^{p} \left(\frac{\lambda_{n,\max}}{\lambda_{n,\max} + \nu_{z_j}}\right)^{1/2} \exp\left\{-\frac{\hat{\phi}_j^2}{2\left(\lambda_{n,\max} + \nu_{z_j}\right)}\right\} \equiv U_f(z).$$

Hence, we have

$$\frac{f(\hat{\phi}|z)}{f(\hat{\phi}|z_o)} \leq \frac{U_f(z)}{L_f(z_o)}. \tag{3.34}$$

We now claim $\frac{U_f(z)}{L_f(z_o)} \xrightarrow{p} 0$ as $n \to 0$ for any $z \neq z_o$. Define

$$H_n(z_j, z_{o,j}) = \left\{\frac{\lambda_{n,\max}(\lambda_{n,\min} + \nu_{z_{o,j}})}{\lambda_{n,\min}(\lambda_{n,\max} + \nu_{z_j})}\right\}^{1/2} \exp\left\{-\frac{\hat{\phi}_j^2}{2(\lambda_{n,\max} + \nu_{z_j})} + \frac{\hat{\phi}_j^2}{2(\lambda_{n,\min} + \nu_{z_{o,j}})}\right\}.$$

Suppose $z_{o,j} = 0$. Then we have that $\hat{\phi}_j = O_p(n^{-1/2})$. Recall that from (**A5**), $\nu_0 = o(n^{-1})$. If $z_j = 0$, then

$$\begin{aligned} H_n(0,0) &= \left\{\frac{\lambda_{n,\max}(\lambda_{n,\min} + \nu_0)}{\lambda_{n,\min}(\lambda_{n,\max} + \nu_0)}\right\}^{1/2} \exp\left\{-\frac{\hat{\phi}_j^2}{2(\lambda_{n,\max} + \nu_0)} + \frac{\hat{\phi}_j^2}{2(\lambda_{n,\min} + \nu_0)}\right\} \\ &= \left\{\frac{O(n^{-2}) + o(n^{-2})}{O(n^{-2}) + o(n^{-2})}\right\}^{1/2} \exp\left\{-\frac{O_p(n^{-1})}{O(n^{-1}) + o(n^{-1})} + \frac{O_p(n^{-1})}{O(n^{-1}) + o(n^{-1})}\right\}. \end{aligned}$$

This implies that $H_n(0,0) = O_p(1)$. From (**A5**), we have $\nu_1 = O(1)$. If $z_j = 1$, then

$$\begin{aligned} H_n(1,0) &= \left\{\frac{\lambda_{n,\max}(\lambda_{n,\min} + \nu_0)}{\lambda_{n,\min}(\lambda_{n,\max} + \nu_1)}\right\}^{1/2} \exp\left\{-\frac{\hat{\phi}_j^2}{2(\lambda_{n,\max} + \nu_1)} + \frac{\hat{\phi}_j^2}{2(\lambda_{n,\min} + \nu_0)}\right\} \\ &= \left\{\frac{O(n^{-2}) + o(n^{-2})}{O(n^{-2}) + O(n^{-1})}\right\}^{1/2} \exp\left\{-\frac{O_p(n^{-1})}{2\{O(n^{-1}) + O(1)\}} + \frac{O_p(n^{-1})}{2\{O(n^{-1}) + o(n^{-1})\}}\right\}. \end{aligned}$$

This implies that $H_n(1,0) = o_p(1)$. Suppose $z_{o,j} = 1$. Then we have $\hat{\phi}_j = O_p(1)$. If $z_j = 0$, then

$$\begin{aligned} H_n(0,1) &= \left\{ \frac{\lambda_{n,\max}(\lambda_{n,\min} + \nu_1)}{\lambda_{n,\min}(\lambda_{n,\max} + \nu_0)} \right\}^{1/2} \exp\left\{ -\frac{\hat{\phi}_j^2}{2(\lambda_{n,\max} + \nu_0)} + \frac{\hat{\phi}_j^2}{2(\lambda_{n,\min} + \nu_1)} \right\} \\ &= \left\{ \frac{O(n^{-2}) + O(n^{-1})}{O(n^{-2}) + o(n^{-2})} \right\}^{1/2} \exp\left\{ -\frac{O_p(1)}{2\{O(n^{-1}) + o(n^{-1})\}} + \frac{O_p(1)}{2\{O(n^{-1}) + O(1)\}} \right\} \\ &= \{O(n)\}^{1/2} \exp\{-O_p(n)\}. \end{aligned}$$

This implies that $H_n(1,0) = o_p(1)$. When $z_j = 1$, we have

$$\begin{aligned} H_n(1,1) &= \left\{ \frac{\lambda_{n,\max}(\lambda_{n,\min} + \nu_1)}{\lambda_{n,\min}(\lambda_{n,\max} + \nu_1)} \right\}^{1/2} \exp\left\{ -\frac{\hat{\phi}_j^2}{2(\lambda_{n,\max} + \nu_1)} + \frac{\hat{\phi}_j^2}{2(\lambda_{n,\min} + \nu_1)} \right\} \\ &= \left\{ \frac{O(n^{-2}) + O(n^{-1})}{O(n^{-2}) + O(n^{-1})} \right\}^{1/2} \exp\left\{ -\frac{O_p(1)}{2\{O(n^{-1}) + O(1)\}} + \frac{O_p(1)}{2\{O(n^{-1}) + O(1)\}} \right\}. \end{aligned}$$

This implies that $H_n(1,1) = O_p(1)$. Note that

$$\frac{U_f(z)}{L_f(z_o)} \propto \prod_{j=1}^{p} H_n(z_j, z_{o,j}).$$

If $z \neq z_o$, then $\prod_{j=1}^{p} H_n(z_j, z_{o,j})$ must include at least one of $H_n(1,0)$ or $H_n(0,1)$. This implies that $\prod_{j=1}^{p} H_n(z_j, z_{o,j}) = o_p(1)$ for any $z \neq z_o$. This completes our proof.

Proof of Claim (3.29)

First, we show that our sampling distribution defined in (3.13) converges to the true limiting distribution in (3.12) as $n \to \infty$ in the sense that

$$g(\hat{\phi}|\phi) = \psi(\hat{\phi}|\phi, \hat{V}) = \psi(\hat{\phi}|\phi, V)\{1 + o_p(1)\},$$

where $\hat{V} = n^{-1}\hat{\Sigma}$. In (**A6**), we have

$$\hat{\Sigma} = \Sigma\{1 + o_p(1)\}.$$

Under (**A4**), this implies that

$$|\hat{\Sigma}|^{-1/2} = |\Sigma|^{-1/2}\{1 + o_p(1)\}.$$

Therefore, we have

$$\psi(\hat{\phi}|\phi, \hat{V}) = \frac{1}{(2\pi)^{\frac{p}{2}}|V|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\hat{\phi} - \phi\right)^{\mathrm{T}} V^{-1}\left(\hat{\phi} - \phi\right)\{1 + o_p(1)\} \right]\{1 + o_p(1)\}.$$

To complete the proof, we need to show that

$$\exp\left[-\frac{1}{2}\left(\hat{\phi}-\phi\right)^{\mathrm{T}} V^{-1}\left(\hat{\phi}-\phi\right) o_p(1)\right] = O_p(1). \tag{3.35}$$

From (3.31), we have

$$\frac{n}{2\lambda_{\max}}\|\hat{\phi}-\phi\|^2 \leq \frac{1}{2}\left(\hat{\phi}-\phi\right)^{\mathrm{T}} V^{-1}\left(\hat{\phi}-\phi\right) \leq \frac{n}{2\lambda_{\min}}\|\hat{\phi}-\phi\|^2,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and the largest eigenvalues of Σ, respectively. From the limiting distribution in (3.12), we have $\|\hat{\phi}-\phi\|^2 = O_p(n^{-1})$. This implies our claim in (3.35). Note that

$$\tilde{p}(z_o|x,\delta) = \frac{\int \psi(\hat{\phi}|\phi,V)p(\phi|z_o)p(z_o)d\phi}{\int\int \psi(\hat{\phi}|\phi,V)p(\phi|z)p(z)d\phi dz},$$

and

$$p_g(z_o|x,\delta) = \frac{\int \psi(\hat{\phi}|\phi,\hat{V})p(\phi|z_o)p(z_o)d\phi}{\int\int \psi(\hat{\phi}|\phi,\hat{V})p(\phi|z)p(z)d\phi dz}.$$

Since we have shown that $\psi(\hat{\phi}|\phi,\hat{V}) = \psi(\hat{\phi}|\phi,V)\{1+o_p(1)\}$, we thus obtain

$$|\tilde{p}(z_o|x,\delta) - p_g(z_o|x,\delta)| \xrightarrow{p} 0,$$

as $n \to \infty$.

Bibliography

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Casella, G., Giron, F. J., Martinez, M. L., and Moren, E. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, 37(3):1207–1228.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5):739–747.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35(4):501–514.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press.

Kyung, M., Gilly, J., Ghosh, M., and Casella, G. (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, 5:369–412.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press.

Narisetty, N. N., He, X., et al. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103:681–686.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, pages 581–592.

Sang, H. and Kim, J. K. (2017). An approximate Bayesian inference on propensity score estimation under nonresponse. *Submitted. Available at arXiv:1702.03453*.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 310–313.

Soubeyrand, S. and Haon-Lasportes, E. (2015). Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC. *Statistics & Probability Letters*, 107:84–92.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge university press.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

CHAPTER 4. A PROFILE LIKELIHOOD APPROACH TO SEMIPARAMETRIC ESTIMATION WITH NONIGNORABLE NONRESPONSE

Hejian Sang Kosuke Morikawa Jae Kwang Kim

Abstract

Statistical inference with nonresponse is quite challenging, especially when the response mechanism is not missing at random. The existing methods often require correct model specification for both the outcome regression model and the response model. However, due to nonresponse, both model assumptions cannot be verified from the data and model misspecification can lead to biased inference seriously. To overcome this limitation, we develop a robust semiparametric method based on the profile likelihood obtained from semiparametric response model. The proposed method uses the observed regression model and the semiparametric response model to achieve robustness. An efficient algorithm using fractional imputation is developed. The bootstrap testing procedure is also proposed to test ignorability assumption. The consistency and asymptotic normality of the proposed method are established. The finite-sample performance is examined in the limited simulation studies and an application to the Korean Labor and Income Panel Study dataset is also presented.

key words: Fractional imputation, Kernel regression, Partially generalized linear model, Profile likelihood, Test

4.1 Introduction

Missing data is frequently encountered in statistics. The complete-case method with ignoring missing data can lead to biased estimation and misleading inference (Rubin, 1976; Little and Rubin,

2014). To adjust for the bias due to missing data, some assumption about the response model is often required. If the response probability does not depend on the unobserved variable, the response mechanism is called missing at random (Rubin, 1976). Otherwise, the response mechanism is called not missing at random, also referred to nonignorable missingness. Under the assumption of missing at random, popular statistical tools include propensity score weighting, multiple imputation and fractional imputation. See Rosenbaum and Rubin (1983), Rosenbaum et al. (1987), Rubin (2004) and Kim (2011) for examples. Nonignorable missingness is more challenging than missing at random, since the response model cannot be estimated from the data without extra assumptions. Furthermore, both models cannot be justified from the observed data due to missingness.

Let Y be the study variable that is subject to missingness. Let X be the covariate variable that is always observed. Let δ be the response indicator function of Y, in the sense that $\delta = 1$ if Y is observed, otherwise, $\delta = 0$. Under the assumption of nonignorable nonresponse, Diggle and Kenward (1994) propose a fully parametric method, which assumes parametric models for $f(Y|X)$ and $\text{pr}(\delta = 1 \mid X, Y)$.The fully parametric method is very sensitive to model misspecification. Scharfstein et al. (1999) , Andrea et al. (2001) and Van Dyk and Meng (2012) suggest the sensitivity analysis for the fully parametric method. Instead of assuming the parametric model for $f(Y \mid X)$, Riddles et al. (2016) propose using $f(Y \mid X, \delta = 1)$. Since the data to fit $f(Y \mid X, \delta = 1)$ are fully available, the model assumption about $f(Y \mid X, \delta = 1)$ can be verified from the data. However, it is still a parametric approach subject to model misspecification problem.

To achieve model robustness, Kott and Chang (2010) use a parametric model for $\text{pr}(\delta = 1 \mid X, Y)$ and estimate the parameters by generalized method of moments. This proposed method avoids making the additional assumption on the outcome regression model. The method of Kott and Chang (2010) is still subject to model misspecification of $\text{pr}(\delta = 1 \mid X, Y)$ and is not as efficient as the likelihood method. Furthermore, Morikawa and Kim (2016) propose a semiparametric maximum likelihood method with the parametric assumption on the response model and use the nonparametric kernel method to approximate $f(Y \mid X, \delta = 1)$. Note that all these methods are based on the assumption of correctly specified response model and the model specification can

not be verified. To improve the robustness of the response model, Kim and Yu (2011) consider a semiparametric model. Their proposed method requires validation sample to estimate parameters in the response model. Shao et al. (2016) extend this method to avoid the requirement of validation sample. Both methods assume that response model is the generalized linear function of Y. Under nonignorable nonresponse, we believe that Y plays a critical role in the response model. The generalized linearity assumption of Y in the response model can be limited. We will verify this claim from the simulation study.

All of these issues motivate us to propose a more robust method to handle nonignorable nonresponse. The proposed method uses the generalized partially linear model with nonparametric function of Y. The estimation method is developed from the profile likelihood method. An efficient computation algorithm is proposed based on the fractional imputation (Kim, 2011). Furthermore, hypothesis testing procedure can be developed to test if the response mechanism is missing at random. The proposed method is robust, since the observed regression model can be justified from the data directly and the response mechanism is an unspecified function of Y.

The rest of this paper is organized as follows. The basic setup of nonignorable nonresponse is introduced in §4.2. The proposed method and the computation algorithm is presented in §4.3. In §4.4, the consistency of the proposed method and the asymptotic property are established. The ignorability test is proposed in §4.5. The performance of the proposed method is examined through simulation studies in §4.6. The proposed method is applied to the Korean Labor and Income Panel Study dataset in §4.7. Some discussion and future work are shown in §4.8. Technical proofs are given in Appendix.

4.2 Setup

Suppose that the sample observations $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ are n independent and identically distributed realizations from the random vector (X, Y). Assume x_i are fully observed

and y_i are subject to missingness. Let δ_i be the response indicator function of y_i, in the sense that

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

The parameter of interest is $\theta \in \Theta$, which is uniquely determined from the estimating equation $E\{U(\theta; X, Y)\} = 0$. Under complete data, θ can be estimated by solving

$$\frac{1}{n}\sum_{i=1}^{n} U(\theta; x_i, y_i) = 0. \tag{4.1}$$

However, if nonresponse occurs, the estimating equation in (4.1) cannot be used directly.

Assume that δ_i independently follow a Bernoulli distribution with the success probability $\pi(x_i, y_i)$, where $\pi(x_i, y_i) = \text{pr}(\delta_i = 1|x_i, y_i)$. Then, a consistent estimator of θ could be obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi(x_i, y_i)} U(\theta; x_i, y_i) = 0, \tag{4.2}$$

if $\pi(x_i, y_i)$ were known.

We assume that the response mechanism is not missing at random, in the sense that the response mechanism depends on unobserved y. Under the assumption of not missing at random, we can build the outcome model as $f(y \mid x; \zeta)$ and the response model as $\pi(x, y; \phi)$, where (ζ, ϕ) are unknown parameters. Under fully parametric assumptions, the observed likelihood function is

$$L_{obs}(\phi, \zeta) = \prod_{i=1}^{n} \{\pi(x_i, y_i; \phi) f(y_i|x_i; \zeta)\}^{\delta_i} \left[\int \{1 - \pi(x_i, y; \phi)\} f(y|x_i; \zeta) dy\right]^{1-\delta_i}. \tag{4.3}$$

To avoid the non-identifiability, we also assume that

$$\text{pr}(\delta_i = 1 \mid x_i, y_i) = \text{pr}(\delta_i = 1 \mid x_{i1}, y_i) = \pi(x_{i1}, y_i),$$

where $x_i = (x_{i1}, x_{i2})$ and x_{i2} is the response instrumental variable (Wang et al., 2014). However, the parametric assumptions cannot be justified and the fully parametric method can suffer model misspecification.

Kim and Yu (2011) and Shao et al. (2016) proposed a semiparametric model for the response mechanism. They assume the response model can be expressed as

$$\text{pr}(\delta_i = 1|x_i, y_i) = \frac{\exp\{g(x_{i1}) + \gamma y_i\}}{1 + \exp\{g(x_{i1}) + \gamma y_i\}}, \tag{4.4}$$

where $g(\cdot)$ is unspecified and γ is the tilting parameter that describes the level of nonignorability.

The consistency of the semiparametric estimation in Kim and Yu (2011) and Shao et al. (2016) requires the correct assumption of the response model (4.4). Even though they leave $g(\cdot)$ unspecified, the role of Y in the response model is limited to be generalized linear.

Under the assumption of not missing at random, we believe that the role of Y in the response model is very important. Therefore, we develop an alternative method to model the response mechanism without the generalized linearity assumption of Y. Note that, under assumption (4.4), the predictive model for nonresponse is

$$f(y \mid x, \delta = 0) = f(y \mid x, \delta = 1)\frac{\exp(-\gamma y)}{E\left[\exp(-\gamma y) \mid x, \delta = 1\right]},$$

and the conditional expectation of Y among nonresponse becomes

$$E\left(Y \mid X, \delta = 0\right) = \frac{\int y \exp(-\gamma y) f(y \mid X, \delta = 1) dy}{\int \exp(-\gamma y) f(y \mid X, \delta = 1) dy}.$$

However, such assumption may be unrealistic as the log of nonresponse odd function can only be quadratic functions of Y (Kim and Yu, 2011).

To cover a more general class of nonignorable nonresponse, we assume the response function satisfies

$$Pr(\delta_i = 1 | x_i, y_i) = \frac{\exp\left\{x_{i1}^T \phi + g(y_i)\right\}}{1 + \exp\left\{x_{i1}^T \phi + g(y_i)\right\}}, \tag{4.5}$$

where ϕ is the unknown parameter and $g(\cdot)$ is an unspecified function. Thus, the predictive model for nonresponse is

$$f(y \mid x, \delta = 0) = f(y \mid x, \delta = 1)\frac{\exp\{-g(y)\}}{E\left[\exp\{-g(y)\} \mid x, \delta = 1\right]}. \tag{4.6}$$

Note that $f(y \mid x, \delta = 1)$ can be estimated and validated from the observed data and $g(y)$ is unspecified. Thus, the prediction model (4.6) has less chance to suffer model misspecification. The details of the proposal is presented in next Section.

4.3 Proposed Method

Under the setup in §4.2, we assume the response model satisfies equation (4.5). To avoid the non-identifiable issue between $x_{i1}^T\phi$ and $g(y_i)$, we assume that x_{i1} exclude the intercept. Let

$$\pi\left\{x_{i1}^T\phi + g(y_i)\right\} = \frac{\exp\left\{x_{i1}^T\phi + g(y_i)\right\}}{1+\exp\left\{x_{i1}^T\phi + g(y_i)\right\}}.$$

Thus, if $g(y_i) = \phi_0 + \phi_1 y_i$, the response model turns to the logistic model. Moreover, the response mechanism degenerates to missing at random, if $g(y_i) = \phi_0$. Then, define the nonresponse odds function as

$$O(x_i, y_i) = \frac{\text{pr}(\delta = 0 \mid x_i, y_i)}{\text{pr}(\delta = 1 \mid x_i, y_i)},$$

which leads to $O(x_i, y_i) = \exp\left\{-x_{i1}^T\phi - g(y_i)\right\} = O(\phi, g; x_{i1}, y_i)$ under model assumption (4.5) and the instrumental assumption.

To estimate ϕ and $g(\cdot)$ under complete response, the maximum profile likelihood method can be applied. Under complete data, the log-likelihood function is

$$l(\phi, g) \quad = \sum_{i=1}^{n} \delta_i \log \pi\left\{x_{i1}^T\phi + g(y_i)\right\} + (1-\delta_i)\log\left[1 - \pi\left\{x_{i1}^T\phi + g(y_i)\right\}\right].$$

The maximum profile likelihood method first keeps ϕ fixed and estimate nonparametric function $g(\cdot)$ as $\hat{g}_\phi(\cdot)$. That is, maximizing

$$\tilde{l}(\phi, g) = \sum_{i=1}^{n}\left(\delta_i \log \pi\left\{x_{i1}^T\phi + g(y)\right\} + (1-\delta_i)\log\left[1 - \pi\left\{x_{i1}^T\phi + g(y)\right\}\right]\right) K_h(y_i - y)$$

to obtain $\hat{g}_\phi(y)$, where $K_h(\cdot)$ is the kernel function with bandwidth h. Then, the profile log-likelihood function is

$$l(\phi, \hat{g}_\phi) \quad = \sum_{i=1}^{n} \delta_i \log \pi\left\{x_{i1}^T\phi + \hat{g}_\phi(y_i)\right\} + (1-\delta_i)\log\left[1 - \pi\left\{x_{i1}^T\phi + \hat{g}_\phi(y_i)\right\}\right].$$

Maximizing $l(\phi, \hat{g}_\phi)$ respect to ϕ leads to the consistent estimator $\hat{\phi}$. See Green and Yandell (1985), Tibshirani and Hastie (1987) and Severini and Wong (1992) for the estimation procedures of the generalized partial linear models. The maximum profile likelihood estimator $\hat{\phi}$ converges to the asymptotic normal distribution with rate $n^{-1/2}$.

However, due to nonresponse, the completed log-likelihood is infeasible. Instead, the observed likelihood is used to estimate parameters in missing data problem. Under nonresponse, we can obtain the observed log-likelihood function as

$$l_{obs}(\phi, g) = \sum_{i=1}^{n} \left[\delta_i \log \pi \left\{ x_{i1}^T \phi + g(y_i) \right\} + (1-\delta_i) E \left(\log \left[1 - \pi \left\{ x_{i1}^T \phi + g(y) \right\} \right] \mid x_i, \delta_i = 0 \right) \right]. \quad (4.7)$$

Note that, in the observed log-likelihood function, nonresponse are integrated out by the predictive model $f(y \mid x, \delta = 0)$. The parametric model assumption about $f(y \mid x, \delta = 0)$ is not justifiable due to nonresponse. Thus, we propose to use $f(y \mid x, \delta = 1)$ and the exponential tilting technique (Kim and Yu, 2011) to avoid the parametric model assumption about $f(y \mid x, \delta = 0)$. We can show that

$$f(y \mid x, \delta = 0) \quad = f(y \mid x, \delta = 1) \frac{\exp\{-g(y)\}}{E\left[\exp\{-g(y)\} \mid x, \delta = 1\right]}, \quad (4.8)$$

where the observed outcome model $f(y \mid x, \delta = 1)$ can be validated using the observed data. Assume the parametric model for Y given x and $\delta = 1$ is $f(y \mid x, \delta = 1; \eta)$, which is known up to η. The consistent estimator of η, say $\hat{\eta}$, can obtained by solving

$$\sum_{i=1}^{n} \delta_i s(\eta; x_i, y_i) = 0, \quad (4.9)$$

where $s(\eta; x_i, y_i) = \partial f(y_i \mid x_i, \delta_i = 1; \eta)/\partial \eta$ is the score function of η. Using (4.8), the observed log-likelihood function in (4.7) can be rewritten as

$$\begin{aligned} l_{obs}(\phi, g \mid \hat{\eta}) \quad &= \sum_{i=1}^{n} \delta_i \log \pi \left\{ x_{i1}^T \phi + g(y_i) \right\} \\ &\quad + (1-\delta_i) \frac{E\left(\log \left[1 - \pi \left\{ x_{i1}^T \phi + g(y) \right\} \right] \exp\{-g(y)\} \mid x_i, \delta_i = 1; \hat{\eta} \right)}{E\left[\exp\{-g(y)\} \mid x_i, \delta_i = 1; \hat{\eta}\right]}. \end{aligned}$$

Applying the maximum profile likelihood method to the observed log-likelihood function $l_{obs}(\phi, g \mid \hat{\eta})$ directly is computationally intensive due to the conditional expectation. To solve this issue, we propose to use the fractional imputation method (Kim, 2011) to estimate ϕ and $g(\cdot)$. The proposed algorithm can be described as follows:

I-Step: For sample unit with $\delta_i = 0$, generate y_{ij}^* independently from $f(y \mid x_i, \delta = 1; \hat{\eta})$, where $\hat{\eta}$ is the consistent estimator of η from solving (4.9), for $j = 1, 2, \cdots, M$.

W-Step: Using the current value $g^{(t)}(y)$ of $\hat{g}(y)$, we can assign the fractional weights as

$$w_{ij}^{*(t)} \propto \exp\{-g^{(t)}(y_{ij}^*)\}, \tag{4.10}$$

where $\sum_j w_{ij}^* = 1$.

M-Step: The maximum profile likelihood method can be applied to the approximate observed log-likelihood function

$$\hat{l}_{obs}(\phi, g \mid w^{*(t)}) = \sum_{i=1}^{n} \left(\delta_i \log \pi \left\{ x_{i1}^T \phi + g(y_i) \right\} + (1-\delta_i) \sum_{j=1}^{M} w_{ij}^{*(t)} \log \left[1 - \pi \left\{ x_{i1}^T \phi + g(y_{ij}^*) \right\} \right] \right),$$

where $w^{*(t)}$ is the set of fractional weights. Maximize $\hat{l}_{obs}(\phi, g \mid w^{*(t)})$ using the maximum profile likelihood method to obtain $\phi^{(t+1)}$ and $g^{(t+1)}(\cdot)$.

Repeat *W-Step* and *M-Step* iteratively until convergence is achieved. The fractional weights in (4.10) only depend on $g(\cdot)$. Since $g(\cdot)$ is modeled by a nonparametric method, the proposed method will automatically generate the fractional weights to make

$$\frac{E\left(\log \left[1 - \pi \left\{ x_{i1}^T \phi + g(y) \right\} \right] \exp\{-g(y)\} \mid x_i, \delta_i = 1; \hat{\eta} \right)}{E\left[\exp\{-g(y)\} \mid x_i, \delta_i = 1; \hat{\eta} \right]} \cong \sum_{j=1}^{M} w_{ij}^* \log \left[1 - \pi \left\{ x_{i1}^T \phi + g(y_{ij}^*) \right\} \right]$$

as close as possible. The detail of *M-Step* is implemented in the following Remark.

Remark 4.1. *Note that, in M-step, we need to apply the profile likelihood method to* $\hat{l}_{obs}(\phi, g \mid w^*)$. *The full maximization of* $\hat{l}_{obs}(\phi, g \mid w^*)$ *for each iteration is not necessary. M-step can be implemented by one-step Newton-Raphson algorithm. Define the smoothed log-likelihood function as*

$$\begin{aligned} \tilde{l}_{obs}(\phi, g \mid w^{*(t)}) = & \sum_{i=1}^{n} \left(\delta_i \log \pi \left\{ x_{i1}^T \phi + g(y) \right\} K_h(y_i - y) \right. \\ & \left. + (1-\delta_i) \sum_{j=1}^{M} w_{ij}^{*(t)} \log \left[1 - \pi \left\{ x_{i1}^T \phi + g(y) \right\} \right] K_h(y_{ij}^* - y) \right). \end{aligned} \tag{4.11}$$

The details of M-Step can be described as the following two steps.

Step 1: We can update ϕ *by*

$$\phi^{(t+1)} = \phi^{(t)} - B_t^{-1} A_t,$$

where

$$A_t = \bigtriangledown \hat{l}_{obs}(\phi, \hat{g}_\phi \mid w^{*(t)})\Big|_{\phi=\phi^{(t)}, g=g^{(t)}}$$

is the marginal gradient and

$$B_t = \bigtriangleup \hat{l}_{obs}(\phi, \hat{g}_\phi \mid w^{*(t)})\Big|_{\phi=\phi^{(t)}, g=g^{(t)}}$$

is the Hessian matrix.

Step 2: Update $g(\cdot)$ *by*

$$g^{(t+1)}(y) = g^{(t)}(y) - \frac{G_t(y)}{H_t(y)},$$

where

$$G_t(y) = \bigtriangledown \tilde{l}_{obs}(\phi, g(y) \mid w^{*(t)})\Big|_{\phi=\phi^{(t+1)}, g=g^{(t)}}$$

is a gradient of the smoothed log-likelihood $\tilde{l}_{obs}(\phi, g(y) \mid w^{*(t)})$ *in (4.11) respect to* $g(y)$ *and*

$$H_t(y) = \bigtriangleup \tilde{l}_{obs}(\phi, g(y) \mid w^{*(t)})\Big|_{\phi=\phi^{(t+1)}, g=g^{(t)}}$$

is a Hessian of $\tilde{l}_{obs}(\phi, g(y) \mid w^{*(t)})$ *respect to* $g(y)$

The derivations of the Step 1 and Step 2 are shown in Appendix 4.9.

Once the convergence of the proposed method is achieved, the final estimator of θ, say $\hat{\theta}$, can be obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi\left\{x_{i1}^T\hat{\phi} + \hat{g}(y_i)\right\}} U(\theta; x_i, y_i) = 0. \tag{4.12}$$

Remark 4.2. *Note that, if* Y *is discrete, then the proposed method is degenerated to the parametric model. For example,* $Y \in \{0, 1\}$*. Then, the response mechanism is*

$$\text{pr}(\delta = 1 \mid x, y) = \frac{\exp\left\{x_1^T\phi + g(y)\right\}}{1 + \exp\left\{x_1^T\phi + g(y)\right\}}, \tag{4.13}$$

which is a parametric function of $\{\phi, g(0), g(1)\}$*.*

Remark 4.3. *It is worth to mentioning that the parametric observed regression model* $f(y \mid x, \delta = 1; \eta)$ *can be build into the fully nonparametric regression model. we can show that for function* $A(\delta, x_1, Y) = \log\{1 - \pi(\phi, g; x_1, Y)\}$*, we have*

$$E\{A(\delta, x_1, Y) \mid x, \delta = 0\} = \frac{\int A(\delta, x_1, y) O(\phi, g; x_1, y) f(y \mid x, \delta = 1) dy}{\int O(\phi, g; x_1, y) f(y \mid x, \delta = 1) dy}.$$

Using the kernel smoothing method, we can approximate $E\{A(\delta, x_1, Y) \mid x, \delta = 0\}$ *as*

$$\hat{E}\{A(\delta, x_1, Y) \mid x, \delta = 0\} = \frac{\sum_{j=1}^{n} \delta_j K_H(x_j - x) O(\phi, g; x_1, y_j) A(\delta, x_1, y_j)}{\sum_{j=1}^{n} \delta_j K_H(x_j - x) O(\phi, g; x_1, y_j)}. \tag{4.14}$$

Since we have already shown that $O(\phi, g; x_1, y) = \exp\left\{-\phi^T x_1 - g(y)\right\}$*, we can simply (4.14) as*

$$\hat{E}\{A(\delta, x_1, Y) \mid x, \delta = 0\} = \frac{\sum_{j=1}^{n} \delta_j K_H(x_j - x) \exp\{-g(y_j)\} A(\delta, x_1, y_j)}{\sum_{j=1}^{n} \delta_j K_H(x_j - x) \exp\{-g(y_j)\}}. \tag{4.15}$$

Using (4.15) to replace the conditional expectation in $l_{obs}(\phi, g \mid \hat{\eta})$, we can build the observed log-likelihood function without parametric assumption about $f(y \mid x, \delta = 1)$.

4.4 Asymptotic Theory

In this section, we establish the consistency and the asymptotic normality of the proposed estimator in (4.12). The following assumptions are sufficient conditions.

C1: The true response model $\pi(x, y)$ satisfies (4.5).

C2: The kernel function $K(\cdot)$ satisfies the following properties

$K(u) = 0$ for $|u| > 1$;

$\sup_u |K(u)| < \infty$;

$\int K(u)du = 1, \int uK(u)du = 0, \int u^2 K(u) < \infty$.

The bandwidth h satisfies $h \to 0$ and $nh \to \infty$.

C3: Regularity conditions to establish the asymptotic normality of $\hat{\eta}$.

C4: Regularity conditions for the partially logistic linear models, including *Assumptions 1–4* in Appendix 4.11.

C5: Regularity conditions for estimating equation (4.12).

Condition (**C1**) is our semiparametric model assumption and we will test robustness of our proposed method to this assumption in numerical studies. (**C2**) is a standard assumption for kernel method. The regularity conditions in (**C3**) are standard conditions to obtain asymptotic normality of maximum likelihood estimator $\hat{\eta}$. (**C4**) introduces the sufficient conditions to establish the asymptotic normality of $\hat{\phi}$ under complete data. (**C5**) includes the regularity conditions for estimating equation (4.12). The details of (**C3**)and (**C5**) are shown in Appendix 4.11.

Lemma 4.1. *Under Conditions **C1–C4**, our proposed algorithm enjoys the monotone increasing property, in the sense of*

$$\hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)}) \leq \hat{l}_{obs}(\phi^{(t+1)}, g_{\phi^{(t+1)}} \mid w^{*(t)}), \tag{4.16}$$

$$\tilde{l}_{obs}(\phi^{(t+1)}, g^{(t)} \mid w^{*(t)}) \leq \tilde{l}_{obs}(\phi^{(t+1)}, g^{(t+1)} \mid w^{*(t)}), \tag{4.17}$$

where $\tilde{l}(\phi, g \mid w^{*(t)})$ is defined in (4.11) for any y.

The proof of Lemma 4.1 is shown in Appendix 4.11. From Lemma (4.1), the estimators from our proposed algorithm make the marginal observed log-likelihood of ϕ and the smoothed log-likelihood of g keep increasing.

Theorem 4.1. *Under conditions **C1–C4**, we establish that*

$$\sqrt{n}(\hat{\phi} - \phi_0) \to N(0, \Sigma_1), \tag{4.18}$$

in distribution, as $n, M \to \infty$. ϕ_0 is the true parameter value and $\Sigma_1 = \tilde{I}_{obs}^{-1} + \Sigma_2 + \Sigma_3$. $\tilde{I}_{obs}$ is the observed Fisher information. Σ_2 is the variability of estimating η_0 and Σ_3 is the covariance between $\hat{\phi}$ and $\hat{\eta}$.

The proof of Theorem 4.1 is presented in Appendix 4.11. From Theorem 4.1, we can see that our proposed method has $\sqrt{n}$ convergence rate for parameters, which is the same for fully parametric models.

Theorem 4.2. *Under conditions **C1**–**C5**, we have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, \Sigma), \tag{4.19}$$

where θ_0 is the true value and $\Sigma > 0$.

The proof of Theorem 4.2 is shown in Appendix 4.12. In Appendix 4.12, we have

$$\hat{\theta} - \theta_0 \cong -\left[E\left\{\frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \theta_0}\right\}\right]^{-1}\left[U(\theta_0 \mid \phi_0, g_0) + E\left\{\frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \phi_0}\right\}(\hat{\phi} - \phi_0)\right].$$

Then, we can see that Σ is composite of variability from estimating equation (4.12) and variability from estimating ϕ_0, which already covers estimating η_0.

4.5 Ignorability Test

In §4.2, we assume the response mechanism satisfies (4.5). Thus, if $g(y)$ is a constant, say $g(y) = c$ for some $c \in \mathbb{R}$, the response mechanism degenerates to missing at random. If we are confident that the response mechanism is missing at random, estimation and inference can be greatly simplified without worrying about nonignorable bias. Our response model is a semiparametric model of y. It is a great interest to test if the response mechanism is missing at random.

Under the null hypothesis $\mathrm{H}_0 : g(y) = c$, the response mechanism is a parametric model of unknown (ϕ, c). Thus, (ϕ, c) can be estimated from maximizing the log-likelihood function. That is to maximize

$$l(\phi, c) = \sum_{i=1}^{n} \delta_i \log \pi(\phi, c; x_i) + (1 - \delta_i) \log\left\{1 - \pi(\phi, c; x_i)\right\}, \tag{4.20}$$

respect to (ϕ, c), where

$$\pi(\phi, c; x_i) = \frac{\exp(x_{i1}^T \phi + c)}{1 + \exp(x_{i1}^T \phi + c)}.$$

The likelihood ratio test statistic does not work here due to the non-negligible smoothing bias and different likelihood functions (smoothed and unsmoothed functions). See Härdle et al. (1998) and Lombardía and Sperlich (2008) for related clarification. To solve this issue, Härdle et al. (1998)

proposed using the weighted distance test statistic based on the quasi-likelihood of logistic model. Under complete response, we propose using

$$R = \sum_{i=1}^{n} \pi(\hat{\phi}_a, \hat{c}_a; x_i) \left\{1 - \pi(\hat{\phi}_a, \hat{c}_a; x_i)\right\} \left\{x_{i1}^T(\hat{\phi} - \hat{\phi}_a) + \hat{g}(y_i) - \hat{c}_a\right\}^2, \tag{4.21}$$

where $(\hat{\phi}_a, \hat{c}_a)$ is the solution of (4.20) and $\hat{\phi}$ is the estimator of the proposed profile method. Under the null hypothesis and some regularity conditions, Härdle et al. (1998) showed

$$v_n^{-1}(R - e_n) \to N(0, 1),$$

in distribution, where (v_n, e_n) is very difficult to compute.

Under nonresponse, the test statistic in (4.21) can be approximated by

$$\begin{aligned} \hat{R} = \quad & \sum_{i=1}^{n} \pi(\hat{\phi}_a, \hat{c}_a; x_i) \left\{1 - \pi(\hat{\phi}_a, \hat{c}_a; x_i)\right\} \left[\delta_i \left\{x_{i1}^T(\hat{\phi} - \hat{\phi}_a) + \hat{g}(y_i) - \hat{c}_a\right\}^2 \right. \\ & \left. + (1 - \delta_i) \sum_{j=1}^{M} w_{ij}^* \left\{x_{i1}^T(\hat{\phi} - \hat{\phi}_a) + \hat{g}(y_{ij}^*) - \hat{c}_a\right\}^2\right]. \end{aligned} \tag{4.22}$$

Remark 4.4. *Note that, under the null hypothesis,*

$$\sum_{j=1}^{M} w_{ij}^* \left\{x_{i1}^T(\hat{\phi} - \hat{\phi}_a) + \hat{g}(y_{ij}^*) - \hat{c}_a\right\}^2 - \left[x_{i1}^T(\hat{\phi} - \hat{\phi}_a) + E\left\{\hat{g}(y) \mid \hat{\eta}, \delta = 1, x_i\right\} - \hat{c}_a\right]^2 \to 0,$$

almost surely, as $M \to \infty$. Thus, we can rewrite

$$\hat{R} = R + \sum_{i=1}^{n} \pi(\hat{\phi}_a, \hat{c}_a; x_i) \left\{1 - \pi(\hat{\phi}_a, \hat{c}_a; x_i)\right\} (1 - \delta_i) \left[\hat{g}(y_i) - E\left\{\hat{g}(y) \mid \hat{\eta}, \delta = 1, x_i\right\}\right]^2.$$

Under the null hypothesis, $E\left[\hat{g}(y_i) - E\left\{\hat{g}(y) \mid \hat{\eta}, \delta = 1, x_i\right\}\right]^2 = o_p(1)$. Thus, $\hat{R} = R\{1 + o_p(1)\}$. We can conclude that $v_n^{-1}(\hat{R} - e_n)$ also converges to the normal distribution. If M is finite, v_n would be inflated by the variability of imputation and $\hat{\eta}$.

Since (v_n, e_n) is unknown and the effect of imputation needs to incorporated, the bootstrap method can be used to test $\mathrm{H_a} : g(y) = c$. Under $\mathrm{H_0} : g(y) = c$, the parametric bootstrap is developed. The algorithm of the parametric bootstrap is shown in Appendix 4.10.

4.6 Simulation Study

4.6.1 Simulation Study I

In this simulation study, we investigate the performance of the proposed method in the finite sample. The robustness of the proposed method is also examined when the response model assumption is violated. The simulation study can be described as a 3×9 factorial design, where the factors are the outcome regression model and the response mechanism. Assume the covariate $x_i = (x_{i1}, x_{i2})$ are generated from $N(u, \Sigma)$ with $u = (1, 1)^T$ and $\Sigma = \text{Diag}(0.25, 0.25)$ independently. For the outcome regression model, let $y_i = m(x_i) + e_i$, where the mean functions $m(x)$ are one of followings:

$$\begin{aligned}
\mathcal{M}_1 : &\quad m(x) = -1 + (x_2 - 0.5)^2 \\
\mathcal{M}_2 : &\quad m(x) = -2.75 + x_1 + x_2 + x_1 x_2 \\
\mathcal{M}_3 : &\quad m(x) = -1.75 + x_1 + x_2
\end{aligned}$$

and $e \sim N(0, 0.25)$.

For the response mechanism, let δ_i be generated from a Bernoulli distribution with the success probability π_i independently. For the true response mechanism, we consider the following setups:

$\mathcal{R}_1$: (Linear MAR)

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_{i1})}{1 + \exp(\phi_0 + \phi_1 x_{i1})},$$

where $(\phi_0, \phi_1) = (0.7, 0.2)$.

$\mathcal{R}_2$: (Linear NMAR)

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i)},$$

where $(\phi_0, \phi_1) = (1, 0.2, 0.2)$.

$\mathcal{R}_3$: (Non-linear NMAR with quadratic term in y)

$$\pi_i = \frac{\exp\left(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i^2\right)}{1 + \exp\left(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i^2\right)},$$

where $(\phi_0, \phi_2, \phi_2) = (0, 0.1, 0.7)$.

$\mathcal{R}_4$: (Non-linear NMAR with quadratic term in both x and y)

$$\pi_i = \frac{\exp\left\{\phi_0 + \phi_1 x_{i1}^2 + \phi_2 y_i^2\right\}}{1 + \exp\left\{\phi_0 + \phi_1 x_{i1}^2 + \phi_2 y_i^2\right\}},$$

where $(\phi_0, \phi_1, \phi_2) = (0, 0.1, 0.5)$.

$\mathcal{R}_5$: (Non-linear NMAR with exponential term in x_1 and quadratic term in y)

$$\pi_i = \frac{\exp\left\{\phi_0 + \phi_1 \exp(x_{i1} - 1) + \phi_2 y_i^2\right\}}{1 + \exp\left\{\phi_0 + \phi_1 \exp(x_{i1} - 1) + \phi_2 y_i^2\right\}},$$

where $(\phi_0, \phi_1, \phi_2) = (0, 0.1, 0.6)$

$\mathcal{R}_6$: (Non-linear NMAR with exponential term in y and interaction term)

$$\pi_i = \frac{\exp\left\{\phi_0 + \phi_1 x_{i1} y_i + \phi_2 y_i^2\right\}}{1 + \exp\left\{\phi_0 + \phi_1 x_{i1} y_i + \phi_2 y_i^2\right\}},$$

where $(\phi_0, \phi_1, \phi_2) = (0, 0.1, 0.6)$.

$\mathcal{R}_7$: (Probit NMAR)

$$\pi_i = \Phi(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i^2),$$

where $(\phi_0, \phi_1, \phi_2) = (0, -0.1, 0.6)$ and $\Phi(\cdot)$ is the normal cumulative distribution function.

$\mathcal{R}_8$: (Complementary log-log NMAR)

$$\pi_i = 1 - \exp\left\{-\exp(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i^2)\right\},$$

where $(\phi_0, \phi_1, \phi_2) = (0, -0.05, 0.3)$.

$\mathcal{R}_9$: (x_1 instrumental variable)

$$\pi_i = \frac{\exp\left(\phi_0 + \phi_1 x_{i2} + \phi_2 y_i^2\right)}{1 + \exp\left(\phi_0 + \phi_1 x_{i2} + \phi_2 y_i^2\right)},$$

where $(\phi_0, \phi_1, \phi_2) = (0, 0.1, 0.7)$.

The response mechanism $\mathcal{R}_1$ is missing at random, in the sense of $g(y) = \phi_0$. $\mathcal{R}_2$ is the logistic linear model assumption, which is mostly used to fit the nonresponse model in Kim and Yu (2011) and Shao et al. (2016). $\mathcal{R}_3$ satisfies all model assumptions of the proposed method. $\mathcal{R}_4$ and $\mathcal{R}_5$ violate the linearity assumption of x_{i1} and $\mathcal{R}_6$ has the interaction term of x_i, y_i, which leads to failure of the linearity assumption. $\mathcal{R}_7$ and $\mathcal{R}_8$ are used to check the robustness of the link function. $\mathcal{R}_9$ is used to check the violation of the instrumental variable assumption.

For each response mechanism, the overall response rates are approximately 70%. For each setup, we generate a Monte Carlo sample with $n = 500$ independently for replication $B = 2,000$. Suppose we are interested in $\theta = E(y)$. Thus, $U(\theta; \boldsymbol{x}, y) = y - \theta$. For each realized sample, we apply the following methods.

1. Full estimator θ_{full}: Use the full sample to estimate θ, but which is not practical in real data analysis.

2. CC estimator θ_{CC}: Ignore nonresponse and only use responses to estimate θ.

3. Kott and Chang (2010)'s method θ_{KC}: Assume the response model is

$$Pr(\delta_i = 1 \mid x_i, y_i) = \pi(\phi; y_i) = \frac{\exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}. \tag{4.23}$$

   And the estimates can be obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi(\phi; x_{1i}, y_i)} - 1\right\}(1, \boldsymbol{x}_i)' = \boldsymbol{0},$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\phi; x_{1i}, y_i)}(y_i - \theta) = 0.$$

4. Riddles et al. (2016)'s method θ_{FI}: The observed regression model is

$$y_i \mid \boldsymbol{x}_i, \delta_i = 1 \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2}, \sigma^2). \tag{4.24}$$

   The response working model uses (4.23).

5. θ_{SP}: The proposed method with x_2 as the response instrumental variable. The bandwidths are chosen by rule of thumb (Silverman, 1986). The working observed regression model is specified as $y_i \mid \boldsymbol{x}_i, \delta_i = 1 \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2}, \sigma^2)$.

The simulation results for $\mathcal{R}_1$ – $\mathcal{R}_3$, $\mathcal{R}_4$ – $\mathcal{R}_6$ and $\mathcal{R}_7$ – $\mathcal{R}_9$ are presented in Table 4.1, 4.2 and 4.3, separately.

Table 4.1: Simulation results (part I) from $B = 2,000$ Monte Carlo studies

Res	Model	Estimates	θ_{full}	θ_{CC}	θ_{KC}	θ_{FI}	θ_{SP}
R_1	M_1	bias	-0.001	-0.002	-0.003	-0.002	-0.005
		std	0.035	0.042	0.045	0.041	0.039
		rmse	0.035	0.042	0.045	0.041	0.039
	M_2	bias	0.001	0.030	0.001	0.001	-0.000
		std	0.067	0.080	0.070	0.069	0.070
		rmse	0.067	0.085	0.070	0.069	0.070
	M_3	bias	0.000	0.015	0.000	0.000	0.000
		std	0.038	0.045	0.044	0.044	0.042
		rmse	0.038	0.048	0.044	0.044	0.042
R_2	M_1	bias	0.001	0.027	-0.000	-0.000	0.003
		std	0.035	0.041	0.043	0.039	0.039
		rmse	0.035	0.049	0.043	0.039	0.039
	M_2	bias	-0.002	0.119	-0.002	-0.002	0.010
		std	0.069	0.080	0.071	0.070	0.071
		rmse	0.069	0.143	0.071	0.070	0.072
	M_3	bias	-0.000	0.045	-0.001	-0.001	0.008
		std	0.039	0.044	0.042	0.043	0.042
		rmse	0.039	0.063	0.042	0.042	0.043
R_3	M_1	bias	0.000	0.098	-0.032	-0.062	-0.004
		std	0.036	0.051	0.053	0.045	0.044
		rmse	0.036	0.110	0.062	0.076	0.044
	M_2	bias	-0.001	0.095	-0.016	-0.036	-0.004
		std	0.068	0.090	0.071	0.069	0.071
		rmse	0.068	0.130	0.073	0.078	0.071
	M_3	bias	-0.001	0.065	-0.001	-0.010	0.006
		std	0.038	0.053	0.045	0.047	0.045
		rmse	0.038	0.084	0.045	0.048	0.045

From Table 4.1, when the response model is logistic linear ($\mathcal{R}_1/\mathcal{R}_2$), all methods are consistent. For quadratic model M_1, θ_{FI} and θ_{SP} are more efficient than θ_{KC}. Under M_2, M_3, θ_{FI} and θ_{SP} are no worse than θ_{KC}. When the response model is logistic quadratic ($\mathcal{R}_3$), θ_{KC} and θ_{FI} are biased under M_1. However, the proposed θ_{SP} is still consistent and has smaller mean square error. When the outcome regression model is M_2, which is slightly violated the linearity, θ_{FI} is biased and θ_{KC}

is slightly biased. The proposed θ_{SP} performs better than θ_{FI} and θ_{KC} in terms of mean square error. When the outcome regression model is linear M_3, θ_{SP} and θ_{KC} are consistent, but θ_{FI} is slightly biased. In terms of efficiency, θ_{SP} and θ_{GMM} are better, because $f(Y \mid X, \delta = 1)$ uses the full models and induces additional noise from the quadratic terms.

From table 4.2, when the linearity assumption of X in response model is violated, the proposed method still works well. For nonlinear outcome regression models (M_1/M_2), θ_{KC} and θ_{FI} are biased due to model misspecification. However, the proposed method is always consistent. For linear outcome regression model (M_3), θ_{KC} and θ_{SP} are consistent.

From Table 4.3, the misspecification of link function in the response model does not effect the consistency of the proposed method. Furthermore, the violation of the instrumental assumption also does not effect the proposed method heavily. In summary, the proposed method outperforms θ_{KC} and θ_{FI}. Also, the proposed method suffers less model misspecification.

4.6.2 Simulation Study II

In this section, we perform simulation studies to validate the proposed test statistic in §4.5. The power of the proposed test is related to the non-constant effect of $g(y)$ and sample size. Thus, we design a 4×2 factorial studies, where factors are the coefficient of $g(y)$ and the sample size.

Assume the superpopulation model is generated as as follows: First, covariate variables $x_i = (x_{i1}, x_{i2})$ are generated independently from multivariate normal distribution with mean $(1, 1)$ and variance $\text{Diag}(0.25, 0.25)$. Second, response variables y_i are generated independently from normal distribution $N(-1 + x_{i1} + x_{i2}, 0.25)$.

Assume the response function is

$$p_i = \frac{\exp(0.1x_{i1} + \phi_y y_i^2)}{1 + \exp(0.1x_{i1} + \phi_y y_i^2)}.$$

The response indicator functions are generated from a simple random sampling with replacement process with approximate response rate being 70%. The first order inclusion probabilities are $\{p_i\}_{i=1}^n$.

The whole simulation process can be described as follows:

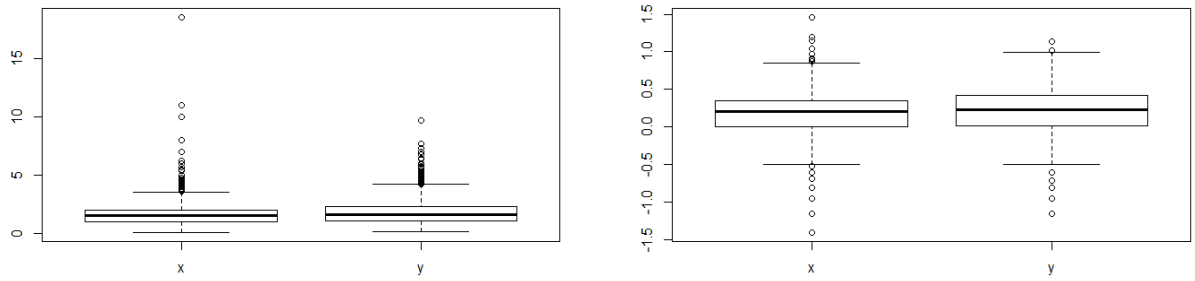
1. Generate the complete sample from the superpopulation model with size $n \in \{100, 500\}$.
2. Apply the response mechanism to create nonresponse with $\{0, 0.2, 0.5, 1\}$.
3. Apply the proposed bootstrap method in Appendix 4.10 to obtain the empirical distribution of the proposed test statistic.
4. Repeat step 1–3 $B = 1,000$ times.

The simulation results are presented in Table 4.4.

The power of the test is that the probability of rejecting the null hypothesis, given that the alternative hypothesis is true. From Table 4.4, the power of the proposed test statistic is increasing as the violation (ϕ_y) of constant $g(y)$ increases for fixed sample size. For fixed ϕ_y, the power of the proposed test statistic also increases as sample size increases. For $\phi_y = 0$, which indicates the null hypothesis is true, the proposed test statistic can achieves the type I error bound approximately when sample size is 500. In summary, the proposed test statistic and the bootstrap method can be used to test the ignorability effectively.

4.7 Application

In this section, the proposed method is applied to Korea Labor and Income Panel Survey (KLIPS). The introduction of the penal survey can be checked out at http://www.kli.re.kr/klips/en/about/introduce.jsp. The study variable (y) is the average monthly income for the current year and the auxiliary variable (x) is the average monthly income for the previous year. The KLIPS has $n = 2,506$ regular wage earners. And the boxplots for x and y are presented in Figure 4.1. Note that both x, y has outliers which cause challenging to the nonparametric smoothing method. Thus, we take the transformation to both x and y.



(a) The original KLIPS data (b) The transformed KLIPS data: $(x, y) \leftarrow \log(x, y)/2$

Figure 4.1: KLIPS data description ($\times 10^6$ Korean Won).

Since the KLIPS data are completed, we artificially create the missingness and then apply the proposed method to the incomplete data. Assume the true response mechanisms are

$$\begin{aligned}
\mathcal{R}_1: &\quad Pr(\delta = 1 \mid x, y) = \{1 + \exp(-1 + y)\}^{-1}, \\
\mathcal{R}_2: &\quad Pr(\delta = 1 \mid x, y) = [1 + \exp\{-2 + \exp(0.5y)\}]^{-1}, \\
\mathcal{R}_3: &\quad Pr(\delta = 1 \mid x, y) = \begin{cases} 0.7 & \text{if } y < 0.5 \\ 0.4 & \text{otherwise} \end{cases}, \\
\mathcal{R}_4: &\quad Pr(\delta = 1 \mid x, y) = \Phi\{-0.1 + 0.1\exp(0.5y)\}.
\end{aligned}$$

The process is described as following:

1. Use Simple Random Sampling without Replacement (SRSWOR) to obtain n sample units.
2. Apply the response mechanism $\mathcal{R}$ to the sample and get the incomplete sample.
3. Apply the proposed method to the incomplete sample and obtain the parameter estimation.

Let $n = 200$ and replicate the process $B = 2,000$ times. For each realized sample, apply Full, CC, Proposed and GMM method to estimate $\theta = E(y)$. The results are shown in Figure 4.2.

From Figure 4.2, we can see that both proposed and GMM methods achieve consistent estimates and their efficiencies are comparable. CC methods are always biased. The proposed method is

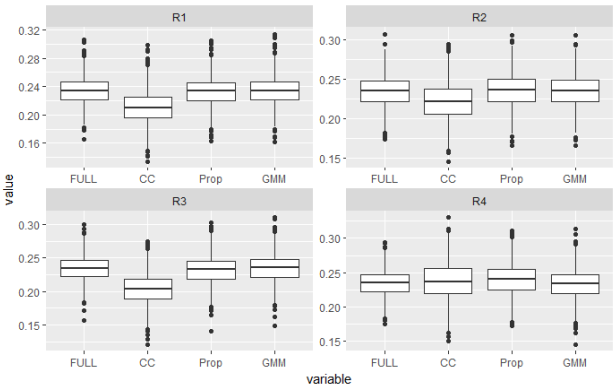


Figure 4.2: Boxpliots of the estimators for Full, CC, Proposed, and GMM methods.

consistent, since it does involve model specifications. The GMM method is consistent in the real data due to the linearity of x and y.

4.8 Discussion

In this paper, we propose a profile likelihood method to achieve robust estimation under a semiparametric nonignorable nonresponse model. From simulation results, our proposed method shows more robustness than generalized linear response models. The proposed method uses the maximum profile likelihood method and an efficient computation algorithm based on fractional imputation is developed. From asymptotic properties, our proposed method enjoys $\sqrt{n}$-consistency. Furthermore, our proposed method assumes the response mechanism is a flexible function of Y. Then, we propose a test procedure to check if the response mechanism is missing at random. The bootstrap method is proposed to obtain the empirical distribution of the proposed test statistic. Our proposed method can be used in survey data directly by replacing the likelihood function to the pseudo likelihood function.

4.9 Appendix A: Derivations in M-Step

Note that, $\tilde{l}_{obs}(\phi, g \mid w^{*(t)})$ are generalized partially linear function of ϕ and g. Then, the profile method likelihood can be applied. The outlined procedures are described as follows. First, $g(y)$

can be estimated by maximizing

$$\begin{aligned}\tilde{l}_{obs}(\phi, g \mid w^{*(t)}) &= \sum_{i=1}^{n} \delta_i \log \pi \left\{ x_{i1}^T \phi + g(y) \right\} K_h(y - y_i) \\ &+ (1-\delta_i) \sum_{j=1}^{M} w_{ij}^{*(t)} \log \left[1 - \pi \left\{ x_{i1}^T \phi + g(y) \right\} \right] K_h(y - y_{ij}^*),\end{aligned}$$

given a fixed ϕ. Denote it as $\hat{g}_\phi(y)$. Then, ϕ can be estimated by maximizing

$$\hat{l}_{obs}(\phi, \hat{g}_\phi \mid w^{*(t)}) = \sum_{i=1}^{n} \delta_i \log \pi \left\{ x_{i1}^T \phi + \hat{g}_\phi(y_i) \right\} + (1-\delta_i) \sum_{j=1}^{M} w_{ij}^{*(t)} \log \left[1 - \pi \left\{ x_{i1}^T \phi + \hat{g}_\phi(y_{ij}^*) \right\} \right].$$

The details of one-step Newton-Raphson algorithm are shown as follows. The maximization of $\tilde{l}_{obs}(\phi, g \mid w^{*(t)})$ respect to $g(y)$ is equivalent to taking the first order derivative respect to $g(y)$. That is

$$\begin{aligned}\frac{\partial \tilde{l}_{obs}(\phi, g \mid w^{*(t)})}{\partial g(y)} &= \sum_{i=1}^{n} \delta_i \left[1 - \pi \left\{ x_{i1}^T \phi + g(y) \right\} \right] K_h(y - y_i) \\ &- (1-\delta_i) \sum_{j=1}^{M} w_{ij}^{*(t)} \pi \left\{ x_{i1}^T \phi + g(y) \right\} K_h(y - y_{ij}^*).\end{aligned}$$

To estimate $g(y)$, it is equivalent to solving $\partial \tilde{l}_{obs}(\phi, g \mid w^{*(t)}) / \partial g(y) = 0$. Applying the one-step Newton-Raphson, we can update the estimator by

$$g(y)^{(t+1)} = g^{(t)}(y) - \frac{G_t(y)}{H_t(y)}$$

where

$$G_t(y) = \sum_{i=1}^{n} \delta_i \left[1 - \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y) \right\} \right] K_h(y - y_i) - (1-\delta_i) \sum_{j=1}^{M} w_{ij}^{*(t)} \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y) \right\} K_h(y - y_{ij}^*)$$

is the gradient of $\tilde{l}_{obs}(\phi, g \mid w^{*(t)})$ respect to $g(y)$, and

$$\begin{aligned}H_t(y) &= -\sum_{i=1}^{n} \left[1 - \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y) \right\} \right] \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y) \right\} \\ &\times \left\{ \delta_i K_h(y - y_i) + (1-\delta_i) \sum_{j}^{M} w_{ij}^{*(t)} K_h(y - y_{ij}^*) \right\},\end{aligned}$$

is the Hessian matrix of $\tilde{l}_{obs}(\phi, g \mid w^{*(t)})$ respect to $g(y)$.

Note that $g(y)$ is the function of ϕ. Thus, take the partial derivative of $\tilde{l}_{obs}(\phi, g \mid w^{*(t)})/\partial g(y)$ respect to ϕ and set it to be 0. That is

$$\begin{aligned}\frac{\partial^2 \tilde{l}_{obs}(\phi,g|w^{*(t)})}{\partial g(y)\partial\phi} &= -\sum_{i=1}^{n}\left[1-\pi\left\{x_{i1}^T\phi+g(y)\right\}\right]\pi\left\{x_{i1}^T\phi+g(y)\right\} \\ &\left\{\delta_i K_h(y-y_i)+(1-\delta_i)\sum_j^M w_{ij}^{*(t)}K_h(y-y_{ij}^*)\right\}\{x_{i1}+\bigtriangledown g(y)\}=0,\end{aligned}$$

where $\bigtriangledown g(y) = \frac{\partial g(y)}{\partial \phi}$. Solving $\partial^2 \tilde{l}_{obs}(\phi, g \mid w^{*(t)})/\{\partial g(y)\partial\phi\} = 0$, we can obtain a closed form for $\bigtriangledown g(y)$ as

$$\bigtriangledown g^{(t)}(y) = \frac{I_t(y)}{H_t(y)},$$

where

$$\begin{aligned}I_t(y) &= \sum_{i=1}^{n}\left[1-\pi\left\{x_{i1}^T\phi^{(t)}+g^{(t)}(y)\right\}\right]\pi\left\{x_{i1}^T\phi^{(t)}+g^{(t)}(y)\right\} \\ &\times\left\{\delta_i K_h(y-y_i)+(1-\delta_i)\sum_j^M w_{ij}^{*(t)}K_h(y-y_{ij}^*)\right\}x_{i1}.\end{aligned}$$

Then, ϕ can be estimated by maximizing

$$\hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)}) = \sum_{i=1}^{n}\delta_i \log\pi\left\{x_{i1}^T\phi+g_\phi(y_i)\right\}+(1-\delta_i)\sum_{j=1}^{M}w_{ij}^{*(t)}\log\left[1-\pi\left\{x_{i1}^T\phi+g_\phi(y_{ij}^*)\right\}\right],$$

which leads to solving

$$\frac{\hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi} = 0.$$

Let

$$\begin{aligned}A_t = \bigtriangledown\hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)}) &= \sum_{i=1}^{n}\delta_i\left[1-\pi\left\{x_{i1}^T\phi^{(t)}+g^{(t)}(y_i)\right\}\right]\left(x_{i1}+\bigtriangledown g^{(t)}(y_i)\right) \\ &-(1-\delta_i)\sum_{j=1}^{M}w_{ij}^{*(t)}\pi\left\{x_{i1}^T\phi^{(t)}+g^{(t)}(y_{ij}^*)\right\}\left(x_{i1}+\bigtriangledown g^{(t)}(y_{ij}^*)\right).\end{aligned}$$

To compute the Hessian matrix of $\hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})$, we consider $\bigtriangledown g$ to be constant with respect to ϕ (Müller, 2001). This leads to

$$\begin{aligned} B_t = \quad & \triangle \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)}) = -\sum_{i=1}^{n} \delta_i \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y_i) \right\} \left[1 - \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y_i) \right\} \right] \\ & \times \left(x_{i1} + \bigtriangledown g^{(t)}(y_i) \right)^{\otimes 2} + (1 - \delta_i) \sum_{j=1}^{M} w_{ij}^{*(t)} \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y_{ij}^*) \right\} \left[1 - \pi \left\{ x_{i1}^T \phi^{(t)} + g^{(t)}(y_{ij}^*) \right\} \right] \\ & \times \left(x_{i1} + \bigtriangledown g^{(t)}(y_{ij}^*) \right)^{\otimes 2}, \end{aligned}$$

where $A^{\otimes 2} = AA^T$. Thus, applying Newton-Raphson algorithm, we can update ϕ by

$$\phi^{(t+1)} = \phi^t - B_t^{-1} A_t.$$

4.10 Appendix B: Algorithm for Bootstrap

From the proposed method in §4.3, a pseudo complete sample $\{(x_i, \hat{y}_i, \delta_i)\}_{i=1}^n$ can be obtained, where

$$\hat{y}_i = \begin{cases} y_i & \text{if } \delta_i = 1 \\ \sum_{j=1}^{M} w_{ij}^* y_{ij}^* & \text{otherwise.} \end{cases}$$

As discussed in §4.5, under the null hypothesis, $(\hat{\phi}_a, \hat{c}_a)$ can be obtained by maximizing (4.20). Then, the proposed parametric bootstrap can be described as follows:

Step 1: Using $(\hat{\phi}_a, \hat{c}_a)$, we can regenerate the response indicators δ_i^* from the Bernoulli distribution with success probability $\pi(\hat{\phi}_a, \hat{c}_a; x_i)$. Then, we can formulate the new pseudo sample $\{x_i, \delta_i^* \hat{y}_i, \delta_i^*\}_{i=1}^n$.

Step 2: Apply $\{x_i, \delta_i^* \hat{y}_i, \delta_i^*\}_{i=1}^n$ to (4.20) to obtain $(\hat{\phi}_a^*, \hat{c}_a^*)$.

Step 3: Apply $\{x_i, \delta_i^* \hat{y}_i, \delta_i^*\}_{i=1}^n$ to the proposed method and compute the test statistic $\hat{R}^k$ in (4.22).

Step 4: Repeat *Step 1–3* B times and compute the p-value as

$$\text{p-value} = \frac{1}{B} \sum_{k=1}^{B} I(\hat{B} < \hat{B}^k).$$

If the p-value is less than the type I error α, then we reject H_0. Otherwise, we have no significant evidence to reject H_0.

4.11 Appendix C: Regularity conditions and Proof of Lemma 4.1 and Theorem 4.1

Regularity conditions of (**C3**) are described as follows.

C3(a): For η in an open subset, assume $s(\eta; X, Y)$ is twice continuously differentiable for every X, Y.

C3(b): Assume there exists η_0, such that $E\{s(\eta_0; X, Y)\} = 0$.

C3(c): For η in a neighborhood of η_0, assume $E\{\|s(\eta; X, Y)\|^2\} < \infty$ and $E\left\{\partial s(\eta; X, Y)/\partial \eta^T\right\}$ exists and is nonsingular.

Regularity conditions of (**C5**) are described as follows.

C5(a): The response probability $\pi(X, Y)$ is bonded below from 0 uniformly.

C5(b): There exists θ_0, such that $E\{U(\theta_0; X, Y)\} = 0$.

C5(c): For θ in a neighborhood of θ_0, assume $U(\theta; X, Y)$ is twice continuously differentiable for every X, Y.

C5(d): For θ in a neighborhood of θ_0, assume $E\{\|U(\theta; X < Y)\|^2\} < \infty$ and $E\left\{\partial U(\theta; X, Y)/\partial \theta^T\right\}$ exists and is nonsingular.

The road map of this proof can be outlined as follows.

Step 1: We will show the asymptotic normality of the profile estimator of β under complete data using

$$l_{Full}(\phi, g) = \sum_{i=1}^{n} \left(\delta_i \log \pi\{\phi; x_{i1}, g(y_i)\} + (1 - \delta_i) \log\left[1 - \pi\{\phi; x_{i1}, g(y_i)\}\right]\right).$$

Step 2: Then, we can establish the asymptotic distribution under nonresponse using

$$\begin{aligned} l_{obs}(\phi, g; \eta_0) = & \sum_{i=1}^{n} \left[\delta_i \log \pi \left\{\phi; x_{i1}, g(y_i)\right\}\right. \\ & \left. +(1-\delta_i) E \left(\log \left[1-\pi \left\{\phi; x_{i1}, g(y)\right\}\right] \mid x_i, \delta_i = 0; \eta_0\right)\right]. \end{aligned}$$

Step 3: The asymptotic distribution is further extended to incorporate the estimation of η_0.

Step 4: Finally, we will show that the proposed algorithm is equivalent to applying the profile method to $l_{obs}(\phi, g; \hat{\eta})$ asymptotically.

Let us first show *Step 1*. Since g maps a scalar y into some space G, define $\zeta = g(y) \in G$. Let

$$p(\delta; \phi, \zeta) = \pi \left\{\phi, \zeta; x_1, y\right\}^{\delta} \left[1-\pi \left\{\phi, \zeta; x_1, y\right\}\right]^{1-\delta}$$

as the conditional distribution of δ given (x, y). Furthermore, let $l(\delta; \phi, \zeta) = \log p(\delta; \phi, \zeta)$. Let $\hat{g}_\phi$ be the solution of maximizing

$$\tilde{l}_{Full}(\phi, g) = \sum_{i=1}^{n} \left(\delta_i \log \pi \left\{\phi; x_{i1}, g(y)\right\} + (1-\delta_i) \log \left[1-\pi \left\{\phi; x_{i1}, g(y)\right\}\right]\right) K_h(y-y_i).$$

Let $\hat{\phi}$ be the maximizer of $l_{Full}(\phi, \hat{g}_\phi)$. Furthermore, we define the Fréchet derivative of $l_{Full}(\phi, g)$ respect to function g as

$$\frac{\partial l_{Full}(\phi, g)}{\partial g} = \left.\frac{\partial l_{Full}(\phi, g+\lambda u)}{\partial \lambda}\right|_{\lambda=0}.$$

Following the proof in Severini and Wong (1992), we present the sufficient conditions to obtain the asymptotic distribution.

Assumption 1. For any fixed $\phi_1 \in \Phi$ and $\zeta_1 \in G$, let

$$\rho(\phi, \zeta) = \int \log p(\delta; \phi, \zeta) p(\delta; \phi_1, \zeta_1) \mathrm{d}\delta.$$

If $\phi \neq \phi_1$, then

$$\rho(\phi, \zeta) < \rho(\phi_1, \zeta_1).$$

Assumption 2. Define the marginal Fisher information for ϕ as

$$\tilde{I}_\phi(\phi,\zeta) = E_{\phi,\zeta}\left\{\frac{\partial l}{\partial \phi}(\delta;\phi,\zeta)^2\right\} - E_{\phi,\zeta}\left\{\frac{\partial l}{\partial \phi}(\delta;\phi,\zeta)\frac{\partial l}{\partial \zeta}(\delta;\phi,\zeta)\right\}^2 E_{\phi,\zeta}\left\{\frac{\partial l}{\partial \zeta}(\delta;\phi,\zeta)^2\right\}^{-1}.$$

Assume $\tilde{I}_\phi(\phi,\zeta) > 0$ for all $\phi \in \Phi$ and $\zeta \in G$.

Assumption 3. Assume that the derivative

$$\frac{\partial^{r+s} l}{\partial \phi^r \partial \zeta^s} l(\delta;\phi,\zeta)$$

exists for all $r \geq 0, s \geq 0, r+s \leq 4$. Moreover,

$$E_0\left\{\sup_\phi \sup_\zeta \left\| \frac{\partial^{r+s} l}{\partial \phi^r \partial \zeta^s} l(\delta;\phi,\zeta)\right\|^2\right\} \leq \infty,$$

where E_0 denotes expectation under the true density function.

Assumption 4. Assume the unction $g(y)$ satisfies the Conditions NP (Nuisance parameter) in Severini and Wong (1992).

The following lemma is established from Severini and Wong (1992) and we are using the special case of logistic semiparametric model.

Lemma 4.2. *Under Assumption 1–4, we can show*

$$\sqrt{n}(\hat{\phi} - \phi_0) \to N(0, \tilde{I}_{\phi_0}^{-1}),$$

where $\tilde{I}_{\phi_0}$ is the marginal Fisher information for ϕ_0. Then, we can also establish that

$$\frac{1}{\sqrt{n}} \frac{d}{d\phi} \frac{\partial l_{Full}(\phi, g_\phi)}{\partial g}\bigg|_{\phi=\phi_0} (\hat{g}_0 - g_0) = o_p(1),$$
$$\frac{1}{\sqrt{n}} \frac{\partial l_{Full}(\phi, g_\phi)}{\partial g}\bigg|_{\phi=\phi_0} (\hat{g}_0' - g_0') = o_p(1),$$

where $g_0 = g_{\phi_0}$ is the true function, $\hat{g}_0 = \hat{g}_{\phi_0}$ and $g' = \frac{dg(y)}{dy}$.

This completes *Step 1. Step 1* is a standard conclusion from Severini and Wong (1992).

Then, we want to extent Lemma 4.2 to nonresponse. Note that $l_{obs}(\phi, g; \eta_0) = E\left\{l_{Full}(\phi, g) \mid X, Y_{obs}, R; \eta_0\right\}$, where $X = (x_1, x_2, \cdots, x_n)$, Y_{obs} is the observed part of $(y_1, \cdots, y_n)$ and $R = (\delta_1, \cdots, \delta_n)$. Similarly, the smoothed observed log-likelihood is $\tilde{l}_{obs}(\phi, g; \eta_0) = E\left\{\tilde{l}_{Full}(\phi, g) \mid X, Y_{obs}, R; \eta_0\right\}$. Then, we can establish the following lemma.

Lemma 4.3. *Let $\hat{g}_\phi$ be the maximizer of $\tilde{l}_{Full}(\phi, g)$, then $\hat{g}_{\phi,obs} = E(\hat{g}_\phi \mid X, Y_{obs}, R; \eta_0)$ is the maximizer of $\tilde{l}_{obs}(\phi, g; \eta_0)$.*

The proof can be briefly shown as follows. We can use the Fréchet derivative and expanse

$$\begin{aligned}\tilde{l}_{Full}(\phi, g) \cong\ & \tilde{l}_{Full}(\phi, \hat{g}_\phi) + \frac{\partial \tilde{l}_{Full}(\phi, g)}{\partial g}\bigg|_{g=\hat{g}_\phi} (g - \hat{g}_\phi) + \frac{\partial^2 \tilde{l}_{Full}(\phi, g)}{\partial g^2}\bigg|_{g=\hat{g}_\phi} (g - \hat{g}_\phi)^2 \\ =\ & \tilde{l}_{Full}(\phi, \hat{g}_\phi) + \frac{\partial^2 \tilde{l}_{Full}(\phi, g)}{\partial g^2}\bigg|_{g=\hat{g}_\phi} (g - \hat{g}_\phi)^2.\end{aligned}$$

Taking the conditional expectation to both sides, we can obtain that

$$\begin{aligned}\tilde{l}_{obs}(\phi, g; \eta_0) \cong\ & E\left\{\tilde{l}_{Full}(\phi, \hat{g}_\phi) \mid X, Y_{obs}, R; \eta_0\right\} \\ & + E\left\{\frac{\partial^2 \tilde{l}_{Full}(\phi, g)}{\partial g^2}\bigg|_{g=\hat{g}_\phi} \mid X, Y_{obs}, R; \eta_0\right\} E\left\{(g - \hat{g}_\phi)^2 \mid X, Y_{obs}, R; \eta_0\right\}.\end{aligned}$$

The above equation is upper-bounded at $\hat{g}_{\phi,obs}$. Then, we complete the proof of Lemma 5.4.

Then, denote $\hat{\phi}_{obs}$ be the solution of maximizing

$$\tilde{l}_{obs}(\phi, \hat{g}_\phi; \eta_0) = E\left\{\tilde{l}_{full}(\phi, \hat{g}_\phi; \eta_0) \mid X, Y_{obs}, R; \eta_0\right\}.$$

Using Lemma 4.2 and following the same procedures in Severini and Wong (1992), we can show that 4.2 also holds for $\tilde{l}_{obs}(\phi, g; \zeta_0)$, in the sense of

Lemma 4.4. *Assume $\inf_{\phi,g,x,y} \pi(\phi, g; x_1, y) > 0$. Under the same assumptions in Lemma 4.2, we can show that show*

$$\sqrt{n}(\hat{\phi}_{obs} - \phi_0) \to N(0, \tilde{I}_{obs}^{-1}),$$

where $\tilde{I}_{obs}$ is the marginal Fisher information for ϕ_0 using the observed log-likelihood function. Then, we can also establish that

$$\frac{1}{\sqrt{n}} \frac{d}{d\phi} \frac{\partial l_{obs}(\phi, g_\phi)}{\partial g}\bigg|_{\phi=\phi_0} (\hat{g}_0 - g_0) = o_p(1).$$

This completes *Step 2*.

Note that $\hat{\phi}_{obs}$ in Lemma 4.3 is a function of η_0 and we can denote it as $\hat{\phi}_{obs}(\eta_0)$. However, our profiled estimation is applied to $\tilde{l}_{obs}(\phi, \hat{g}_{\phi,obs}; \hat{\eta})$, where $\hat{\eta}$ is a solution of

$$U(\eta) = \sum_{i=1}^{n} \delta_i s(\eta; x_i, y_i) = 0.$$

Under the regularity conditions of Z-statistics in Van der Vaart (1998), we can establish that

$$\sqrt{r}(\hat{\eta} - \eta_0) \to N(0, S), \tag{4.25}$$

in distribution, where $r = \sum_{i=1}^{n} \delta_i$ and

$$r \left\{ \frac{\partial U(\eta)}{\partial \eta^T} \right\}^{-1} \text{var}\left\{U(\eta)\right\} \left[\left\{ \frac{\partial U(\eta)}{\partial \eta^T} \right\}^{-1} \right]^T \to S$$

in probability.

To obtain the limiting distribution of $\hat{\phi}_{obs}(\hat{\eta})$, militarization can be used.

$$\hat{\phi}_{obs}(\hat{\eta}) \cong \hat{\phi}_{obs}(\eta_0) + \frac{\hat{\phi}_{obs}(\eta_0)}{\partial \eta_0}(\hat{\eta} - \eta_0).$$

Moreover, $\hat{\phi}_{obs}(\eta_0)$ is the solution of

$$\frac{\partial l_{obs}(\phi, \hat{g}_{\phi,obs}; \eta_0)}{\partial \phi} = 0.$$

Using the derivative of implicit function, we can obtain that

$$\frac{\partial \hat{\phi}(\eta_0)}{\partial \eta_0} = -\left\{ \frac{\partial^2 l_{obs}(\phi, \hat{g}_{\phi,obs}; \eta_0)}{\partial \phi \partial \phi^T} \right\}^{-1} \frac{\partial^2 l_{obs}(\phi, \hat{g}_{\phi,obs}; \eta_0)}{\partial \phi \partial \eta_0^T} \Bigg|_{\phi=\hat{\phi}_{obs}(\eta_0)}.$$

Furthermore,

$$-\left\{ \frac{\partial^2 l_{obs}(\phi, \hat{g}_{\phi,obs}; \eta_0)}{\partial \phi \partial \phi^T} \right\}^{-1} \Bigg|_{\phi=\hat{\phi}_{obs}(\eta_0)} \to n^{-1} \tilde{I}_{obs}^{-1}$$

in probability. Let

$$\hat{C}_n = \frac{\partial^2 l_{obs}(\phi, \hat{g}_{\phi,obs}; \eta_0)}{\partial \phi \partial \eta_0^T} \Bigg|_{\phi=\hat{\phi}_{obs}(\eta_0)} = O_p(n).$$

Thus, we have

$$\hat{\phi}_{obs}(\hat{\eta}) \cong \hat{\phi}_{obs}(\eta_0) + n^{-1} \tilde{I}_{obs}^{-1} \hat{C}_n (\hat{\eta} - \eta_0). \tag{4.26}$$

Combining (4.25) and (4.26), we have

$$\hat{\phi}_{obs}(\hat{\eta}) \to \phi_0, \tag{4.27}$$

in probability, since $\hat{\phi}_{obs}(\eta_0) \to \phi_0$, $n^{-1}\tilde{I}_{obs}^{-1}\hat{C}_n = O_p(1)$ and $\hat{\eta} - \eta_0 = o_p(1)$. Then, we can decompose the variance of $\hat{\phi}_{obs}(\hat{\eta})$ as

$$\begin{aligned} n\text{var}\left\{\hat{\phi}_{obs}(\hat{\eta})\right\} &\cong n\text{var}\left\{\hat{\phi}_{obs}(\eta_0) + n^{-1}\tilde{I}_{obs}^{-1}\hat{C}_n(\hat{\eta} - \eta_0)\right\} \\ &\cong \tilde{I}_{obs}^{-1} + n^{-1}r^{-1}\tilde{I}_{obs}^{-1}\hat{C}_n S\hat{C}_n^T\tilde{I}_{obs}^{-1} + 2n\text{Cov}\left\{\hat{\phi}_{obs}(\eta_0), n^{-1}\tilde{I}_{obs}^{-1}\hat{C}_n(\hat{\eta} - \eta_0)\right\} \\ &\to \tilde{I}_{obs}^{-1} + \Sigma_2 + \Sigma_3, \end{aligned} \tag{4.28}$$

in probability.

Using (4.27) and (4.28), we can show that

$$\sqrt{n}\left\{\hat{\phi}_{obs}(\hat{\eta}) - \eta_0\right\} \to N(0, \Sigma_1), \tag{4.29}$$

where $\Sigma_1 = \tilde{I}_{obs}^{-1} + \Sigma_2 + \Sigma_3$. This completes *Step 3*.

Define

$$\hat{l}_{obs}(\phi, g \mid w^{*(t)}) = \sum_{i=1}^{n}\left(\delta_i \log \pi\left\{\phi; x_{i1}, g(y_i)\right\} + (1-\delta_i)\sum_{j=1}^{M} w_{ij}^{*(t)} \log\left[1 - \pi\left\{\phi; x_{i1}, g(y_{ij}^*)\right\}\right]\right) \tag{4.30}$$

The smoothed function is

$$\begin{aligned} \tilde{l}_{obs}(\phi, g \mid w^{*(t)}) = \textstyle\sum_{i=1}^{n} \quad & \left(\delta_i \log \pi\left\{\phi; x_{i1}, g(y)\right\} K_h(y - y_i)\right. \\ & \left. + (1-\delta_i)\sum_{j=1}^{M} w_{ij}^{*(t)} \log\left[1 - \pi\left\{\phi; x_{i1}, g(y_{ij}^*)\right\}\right] K_h(y - y_{ij}^*)\right). \end{aligned}$$

In our proposed algorithm, *M-Step* is to implement one-step Newton-Raphson method. Finally, we show the following lemma.

Lemma 4.5. *For our proposed algorithm, we have*

$$\begin{aligned} \hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)}) &\le \hat{l}_{obs}(\phi^{(t+1)}, g_{\phi^{(t+1)}} \mid w^{*(t)}), \\ \tilde{l}_{obs}(\phi^{(t+1)}, g^{(t)} \mid w^{*(t)}) &\le \tilde{l}_{obs}(\phi^{(t+1)}, g^{(t+1)} \mid w^{*(t)}). \end{aligned}$$

Given $w^{*(t)}$, the implementation of *M-step* is

$$\phi^{(t+1)} = \phi^{(t)} - \left\{ \frac{\partial^2 \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi\partial\phi^T} \right\}^{-1} \frac{\partial \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi} \Bigg|_{\phi=\phi^{(t)}}, \tag{4.31}$$

$$g^{(t+1)} = g^{(t)} - \left\{ \frac{\partial^2 \tilde{l}_{obs}(\phi^{(t+1)}, g \mid w^{*(t)})}{\partial g^2} \right\}^{-1} \frac{\partial \tilde{l}_{obs}(\phi^{(t+1)}, g \mid w^{*(t)})}{\partial g} \Bigg|_{g=g^{(t)}}. \tag{4.32}$$

Note that,

$$\begin{aligned} \hat{l}_{obs}(\phi^{(t+1)}, g_{\phi^{(t+1)}} \mid w^{*(t)}) = \; & \hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)}) + \frac{\partial \hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)})}{\partial \left(\phi^{(t)}\right)^T} (\phi^{(t+1)} - \phi^{(t)}) \\ & + \frac{1}{2} (\phi^{(t+1)} - \phi^{(t)})^T \frac{\partial^2 \hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)})}{\partial\phi^{(t)}\partial \left(\phi^{(t)}\right)^T} (\phi^{(t+1)} - \phi^{(t)}) \\ & + o_p(\|\phi^{(t+1)} - \phi^{(t)}\|^2). \end{aligned} \tag{4.33}$$

Plugging (4.31) into (4.33), we can obtain

$$\begin{aligned} \hat{l}_{obs}(\phi^{(t+1)}, g_{\phi^{(t+1)}} \mid w^{*(t)}) = \; & \hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)}) \\ & - \frac{1}{2} \frac{\partial \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi^T} \left\{ \frac{\partial^2 \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi\partial\phi^T} \right\}^{-1} \frac{\partial \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi} \Bigg|_{\phi=\phi^{(t)}}. \end{aligned}$$

Since

$$\frac{\partial \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi^T} \left\{ \frac{\partial^2 \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi\partial\phi^T} \right\}^{-1} \frac{\partial \hat{l}_{obs}(\phi, g_\phi \mid w^{*(t)})}{\partial\phi} \Bigg|_{\phi=\phi^{(t)}} \leq 0,$$

we have

$$\hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)}) \leq \hat{l}_{obs}(\phi^{(t+1)}, g_{\phi^{(t+1)}} \mid w^{*(t)}) \tag{4.34}$$

Similarly, we can show

$$\tilde{l}_{obs}(\phi^{(t+1)}, g^{(t)} \mid w^{*(t)}) \leq \tilde{l}_{obs}(\phi^{(t+1)}, g^{(t+1)} \mid w^{*(t)}),$$

using the Fréchet derivative.

By Monotone convergence theorem, we have

$$l_{obs}(\phi, g_\phi; \hat{\eta}) - \hat{l}_{obs}(\phi^{(t)}, g_{\phi^{(t)}} \mid w^{*(t)}) \to 0,$$

$$\tilde{l}_{obs}(\phi, g; \hat{\eta}) - \tilde{l}_{obs}(\phi^{(t+1)}, g^{(t)} \mid w^{*(t)}) \to 0,$$

in probability and for any y, as $t \to \infty, M \to \infty$.

Thus, we conclude that our proposed algorithm provides the same solutions as applying the profile likelihood method to $l_{obs}(\phi, g; \hat{\eta})$ directly. Thus, our proposed estimators enjoy the same asymptotic distributions in (4.29).

4.12 Appendix D: Proof of Theorem 4.2

Let $\hat{\theta}$ is the solution of

$$U(\theta \mid \hat{\phi}, \hat{g}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi \left\{ x_{i1}^T \hat{\phi} + \hat{g}(y_i) \right\}} U(\theta; x_i, y_i) = 0, \tag{4.35}$$

where $(\hat{\phi}, \hat{g})$ is obtained from our proposed method. Note that $\hat{g} = \hat{g}_{\hat{\phi}}$. Then, we apply the Taylor linearization to (4.35) and obtain

$$\begin{aligned} U(\hat{\theta} \mid \hat{\phi}, \hat{g}) \cong \; & U(\theta_0 \mid \phi_0, \hat{g}_0) + \frac{\partial U(\theta_0 \mid \phi_0, \hat{g}_0)}{\partial \theta_0} (\hat{\theta} - \theta_0) \\ & + \frac{\partial U(\theta_0 \mid \phi_0, \hat{g}_0)}{\partial \phi_0} (\hat{\phi} - \phi_0). \end{aligned} \tag{4.36}$$

Moreover, using Fréchet derivative, we have

$$U(\theta_0 \mid \phi_0, \hat{g}_0) \cong U(\theta_0 \mid \phi_0, g_0) + \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial g_0} (\hat{g}_0 - g_0). \tag{4.37}$$

Using (4.36) and (4.37), we get the final expansion as

$$\begin{aligned} U(\hat{\theta} \mid \hat{\phi}, \hat{g}) \cong \; & U(\theta_0 \mid \phi_0, g_0) + \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \theta_0} (\hat{\theta} - \theta_0) \\ & \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \phi_0} (\hat{\phi} - \phi_0) + \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial g_0} (\hat{g}_0 - g_0). \end{aligned}$$

From Lemma (4.4), we have

$$\frac{1}{\sqrt{n}} \frac{d}{d\phi} \frac{\partial l_{obs}(\phi, g_\phi)}{\partial g} \Big|_{\phi=\phi_0} (\hat{g}_0 - g_0) = o_p(1).$$

Assume

$$\sup_y \left| \frac{1}{\sqrt{n}} \frac{d}{d\phi} \frac{\partial l_{obs}(\phi, g_\phi)}{\partial g} \Big|_{\phi=\phi_0} \right| = O_p(\sqrt{n}).$$

Then $\sup_y |(\hat{g}_0 - g_0)| = o_p(n^{-1/2})$.

Assume

$$\sup_y \left| \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial g_0} \right| = O_p(1).$$

Then,

$$\frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial g_0}(\hat{g}_0 - g_0)$$

is negligible. Thus, we have

$$\begin{aligned} U(\hat{\theta} \mid \hat{\phi}, \hat{g}) \cong \quad & U(\theta_0 \mid \phi_0, g_0) + \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \theta_0}(\hat{\theta} - \theta_0) \\ & \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \phi_0}(\hat{\phi} - \phi_0), \end{aligned}$$

which leads to

$$\hat{\theta} - \theta_0 \cong - \left[E \left\{ \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \theta_0} \right\} \right]^{-1} \left[U(\theta_0 \mid \phi_0, g_0) + E \left\{ \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \phi_0} \right\} (\hat{\phi} - \phi_0) \right]. \quad (4.38)$$

Since

$$U(\theta_0 \mid \phi_0, g_0) \to 0$$

$$\hat{\phi} - \phi_0$$

in probability, we can conclude that

$$\hat{\theta} - \theta_0 \to 0 \quad (4.39)$$

in probability.

Using (4.38), we have

$$\begin{aligned} n \left[E \left\{ \tfrac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \theta_0} \right\} \right]^{-1} \quad & \text{var} \left[U(\theta_0 \mid \phi_0, g_0) + E \left\{ \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \phi_0} \right\} (\hat{\phi} - \phi_0) \right] \\ & \times \left[E \left\{ \frac{\partial U(\theta_0 \mid \phi_0, g_0)}{\partial \theta_0} \right\} \right]^{-1} \to \Sigma, \end{aligned}$$

in probability.

Therefore, our final conclusion is that

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \Sigma) \quad (4.40)$$

in distribution.

Bibliography

Andrea, R., Scharfstein, D., Su, T.-L., and Robins, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, 57(1):103–113.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied statistics*, pages 49–93.

Green, P. J. and Yandell, B. S. (1985). Semi-parametric generalized linear models. In *Generalized linear models*, pages 44–55. Springer.

Härdle, W., Mammen, E., and Müller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, 93(444):1461–1474.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1):119–132.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106(493):157–165.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491):1265–1275.

Little, R. J. and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Lombardía, M. J. and Sperlich, S. (2008). Semiparametric inference in generalized mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):913–930.

Morikawa, K. and Kim, J. K. (2016). Semiparametric adaptive estimation with nonignorable nonresponse data. *arXiv preprint arXiv:1612.09207*.

Müller, M. (2001). Estimation and testing in generalized partial linear models–a comparative study. *Statistics and Computing*, 11(4):299–309.

Riddles, M. K., Kim, J. K., and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, page smv047.

Rosenbaum, P. R. et al. (1987). The role of a second control group in an observational study. *Statistical Science*, 2(3):292–306.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *The Annals of statistics*, pages 1768–1802.

Shao, J., Wang, L., et al. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

Van Dyk, D. A. and Meng, X.-L. (2012). The art of data augmentation. *Journal of Computational and Graphical Statistics*.

Wang, S., Shao, J., and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, pages 1097–1116.

Table 4.2: Simulation results (part II) from $B = 2,000$ Monte Carlo studies

Res	Model	Estimates	θ_{full}	θ_{CC}	θ_{KC}	θ_{FI}	θ_{SP}
R_4	M_1	bias	-0.002	0.085	-0.027	-0.051	-0.002
		std	0.035	0.052	0.053	0.045	0.044
		rmse	0.035	0.100	0.060	0.068	0.044
	M_2	bias	0.001	0.112	0.018	-0.038	-0.001
		std	0.068	0.092	0.071	0.069	0.071
		rmse	0.068	0.145	0.073	0.079	0.071
	M_3	bias	-0.002	0.063	-0.002	-0.011	0.004
		std	0.039	0.054	0.046	0.048	0.046
		rmse	0.039	0.083	0.046	0.049	0.046
R_5	M_1	bias	-0.000	0.092	-0.029	-0.055	-0.002
		std	0.036	0.051	0.054	0.045	0.044
		rmse	0.036	0.105	0.061	0.071	0.045
	M_2	bias	0.001	0.102	0.019	-0.035	-0.001
		std	0.065	0.088	0.068	0.066	0.068
		rmse	0.065	0.134	0.071	0.074	0.068
	M_3	bias	-0.001	0.063	-0.001	-0.010	0.007
		std	0.038	0.053	0.046	0.048	0.045
		rmse	0.038	0.082	0.046	0.049	0.046
R_6	M_1	bias	-0.001	0.113	-0.031	-0.061	0.000
		std	0.036	0.054	0.056	0.047	0.045
		rmse	0.036	0.126	0.064	0.077	0.045
	M_2	bias	-0.000	0.125	0.019	-0.044	-0.001
		std	0.067	0.090	0.070	0.068	0.070
		rmse	0.067	0.154	0.072	0.081	0.070
	M_3	bias	0.000	0.080	0.000	-0.011	0.009
		std	0.040	0.056	0.047	0.049	0.046
		rmse	0.040	0.098	0.047	0.050	0.047

Table 4.3: Simulation results (part III) from $B = 2,000$ Monte Carlo studies

Res	Model	Estimates	θ_{full}	θ_{CC}	θ_{KC}	θ_{FI}	θ_{SP}
R_7	M_1	bias	-0.000	0.092	0.020	-0.038	-0.001
		std	0.068	0.091	0.070	0.068	0.071
		rmse	0.068	0.129	0.073	0.078	0.071
	M_2	bias	-0.000	0.092	0.020	-0.038	-0.001
		std	0.068	0.091	0.070	0.068	0.071
		rmse	0.068	0.129	0.073	0.078	0.071
	M_3	bias	-0.000	0.071	-0.001	-0.011	0.009
		std	0.038	0.056	0.046	0.049	0.046
		rmse	0.038	0.090	0.046	0.050	0.047
R_8	M_1	bias	-0.002	0.069	0.012	-0.024	-0.003
		std	0.068	0.086	0.070	0.068	0.070
		rmse	0.068	0.110	0.071	0.072	0.070
	M_2	bias	-0.002	0.069	0.012	-0.024	-0.003
		std	0.068	0.086	0.070	0.068	0.070
		rmse	0.068	0.110	0.071	0.072	0.070
	M_3	bias	-0.001	0.039	-0.001	-0.005	0.005
		std	0.039	0.051	0.045	0.046	0.045
		rmse	0.039	0.064	0.045	0.046	0.045
R_9	M_1	bias	0.002	0.099	0.016	-0.036	-0.001
		std	0.069	0.089	0.071	0.069	0.071
		rmse	0.069	0.133	0.072	0.078	0.071
	M_2	bias	0.002	0.099	0.016	-0.036	-0.001
		std	0.069	0.089	0.071	0.069	0.071
		rmse	0.069	0.133	0.072	0.078	0.071
	M_3	bias	0.000	0.066	-0.009	-0.018	0.002
		std	0.039	0.055	0.046	0.047	0.046
		rmse	0.039	0.086	0.046	0.051	0.046

Table 4.4: Relative number of rejections from $B = 1,000$ Monte Carlo studies. α is the predetermined type I error.

n	ϕ_y	α				
		0.01	0.05	0.1	0.15	0.2
100	0	0	0	0	0	0
	0.2	0.009	0.036	0.071	0.125	0.188
	0.5	0.013	0.062	0.149	0.251	0.341
	1	0.018	0.093	0.229	0.372	0.517
500	0	0.007	0.037	0.079	0.121	0.161
	0.2	0.039	0.135	0.239	0.344	0.423
	0.5	0.177	0.426	0.634	0.800	0.882
	1	0.344	0.705	0.888	0.980	0.995

CHAPTER 5. SEMIPARAMETRIC FRACTIONAL IMPUTATION USING GAUSSIAN MIXTURE MODELS FOR HANDLING MULTIVARIATE MISSING DATA

Hejian Sang Jae Kwang Kim

Abstract

Item nonresponse is frequently encountered in practice. Ignoring missing data can lose efficiency and lead to misleading inference. Fractional imputation is a statistical tool for handling missing data. However, the parametric fractional imputation of Kim (2011) may be subject to bias due to model misspecification. In this paper, we propose a novel semiparametric fractional imputation method using Gaussian mixture model. The proposed method is computationally efficient and leads to robust estimation. The proposed method is further extended to incorporate the categorical auxiliary information. The asymptotic model consistency under missing data is also established. Several numerical studies are performed to check the finite sample performance of the proposed method.

key words: Item nonresponse, Robust estimation, Survey sampling, Variance estimation.

5.1 Introduction

Missing data is frequently encountered in survey sampling, clinical trials and many other areas. Imputation can be used to handle item nonresponse and several imputation methods have been developed in the literature. Rubin (1996) proposed multiple imputation to create multiple complete data sets. Alternatively, fractional imputation (Kim, 2011) makes one complete data with multiple imputed values and corresponding fractional weights. Little and Rubin (2014) and Kim and Shao (2013) provide comprehensive overviews of the methods for handling missing data.

For multivariate missing data with arbitrary missing patterns, imputation methods are developed to preserve the correlation structure in the imputed data. Judkins et al. (2007) proposed an iterative hot deck imputation procedure that is closely related to the data augmentation algorithm of Tanner and Wong (1987) but did not provide variance estimation. Im et al. (2018) developed fractional hot deck imputation for multivariate missing data and the procedure is implemented in Proc SurveyImputae (SAS version 14.2). Other non-hot-deck imputation procedures for multivariate missing data include the multiple imputation approach of Raghunathan et al. (2001) and parametric fractional imputation of Kim (2011). The approaches of Judkins et al. (2007) and Raghunathan et al. (2001) are based on conditionally specified models and the imputation from the conditionally specified model is subject to the model compatibility problem (Chen, 2010). Conditional models for the different missing patterns calculated directly from the observed patterns may not be compatible with each other. The parametric fractional imputation used the joint distribution to create imputed values, but correct specification of the joint model is challenging under missing data. Furthermore, valid inference after multiple imputation requires congeniality and self-efficiency (Meng, 1994), which is not necessary satisfied in many practical problems (Kim et al., 2006; Yang and Kim, 2016b). Fractional imputation does not suffer such problems.

Note that parametric imputation requires correct model specification. Nonparametric imputation methods, such as kernel regression imputation (Cheng, 1994; Wang and Chen, 2009), are robust but may be subject to curse of dimensionality. It is important to develop a unified, robust and efficient imputation method. The proposed semiparametric method fills in this important gap by considering a flexible method for imputation.

In this paper, to achieve robustness against model misspecification, we develop an imputation procedure based on Gaussian mixture models (GMM). GMM is a very flexible model that can be used to handle outliers, heterogeneity and skewness. Lindsay (1995) and McLachlan and Peel (2004) showed that any continuous distribution can be approximated by a finite Gaussian mixture distribution. The proposed method using GMM makes a compromise between efficiency and

robustness. It is semiparametric in the sense that the number of mixture component is chosen automatically from the data. The computation is relatively simple and efficient.

The proposed method is further extended to handle mixed type data including categorical variable. By allowing the proportion vector of mixture component to depend on categorical auxiliary variable, the proposed fractional imputation using GMM can incorporate the observed information in categorical variables and provide a very flexible tool for imputation.

The paper is structured as follows. The setup of the problem is introduced and a short review of fractional imputation are presented in §5.2. In §5.3, the proposed semiparametric method and its algorithm for implementation are introduced. Some asymptotic results are presented in §5.4. In §5.5, the proposed method is further extended to handle mixed type data. Some numerical studies and a real data application are presented to show the performance of the proposed method in §5.6 and §5.7, respectively. In §5.8, we discuss some conclusion and future works. The technical derivations and proof are presented in Appendix.

5.2 Setup

Consider a p-dimensional vector of study variable $Y = (y_1, y_2, \cdots, y_p)$. Suppose that $\{Y_1, Y_2, \cdots, Y_n\}$ are n independent and identically distributed (IID) realizations from the random vector Y. In this paper, we use the upper case to represent vector or matrix and the lower case to denote the elements within vector or matrix. Assume that we are interested in estimating parameter $\theta \in \Theta$, which is defined through $E\{U(\theta; Y)\} = 0$, where $U(\cdot; Y)$ is the estimating function of θ. With no missingness, a consistent estimator of θ can be obtained by the solution to

$$\frac{1}{n}\sum_{i=1}^{n} U(\theta; Y_i) = 0. \tag{5.1}$$

To avoid unnecessary details, we assume that the solotion to (5.1) exists almost everywhere.

However, due to missingness, the estimating equation in (5.1) cannot be applied directly. To formulate the multivariate missingness problem, we further define the response indicator vector

$R = (r_1, r_2, \cdots, r_p)$ for Y as

$$r_j = \begin{cases} 1 & \text{if } y_j \text{ is observed} \\ 0 & \text{otherwise,} \end{cases} \tag{5.2}$$

where $j = 1, 2, \cdots, p$. We assume that the response mechanism is missing at random (MAR) in the sense of Rubin (1976). We decompose $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} represent the observed and missing parts of Y, respectively. Thus, MAR assumption is described as

$$\Pr\{R = (r_1, r_2, \cdots, r_p) \mid Y_{obs}, Y_{mis}\} = \Pr\{R = (r_1, r_2, \cdots, r_p) \mid Y_{obs}\}, \tag{5.3}$$

where any $r_j \in \{0, 1\}$, $j = 1, 2, \cdots, p$.

Under MAR, a consistent estimator of θ can be obtained by solving the following estimating equation:

$$\bar{U}(\theta) = \frac{1}{n}\sum_{i=1}^{n} E\{U(\theta; Y_i) \mid Y_{i,obs}\} = 0, \tag{5.4}$$

where it is understood that $E\{U(\theta; Y_i) \mid Y_{i,obs}\} = U(\theta; Y_i)$ if $Y_{i,obs} = Y_i$. To compute the conditional expectation in (5.4), the parametric fractional imputation (PFI) method of Kim (2011) can be developed. To apply the PFI, we can assume that the random vector Y follows a parametric model $F_0(Y) \in \{F_\zeta(Y) : \zeta \in \Omega\}$. Under MAR, a consistent estimator $\hat{\zeta}$ can be obtained from the observed likelihood. In PFI, M imputed values for $Y_{i,mis}$, say $\left\{Y_{i,mis}^{*(1)}, Y_{i,mis}^{*(2)}, \cdots, Y_{i,mis}^{*(M)}\right\}$ are generated from a proposal distribution with same support of $F_0(Y)$ and are assigned with fractional weights, say $\{w_{i1}^*, w_{i2}^*, \cdots, w_{iM}^*\}$, so that a consistent estimator of θ can be obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i U(\theta; Y_i) + (1-\delta_i)\sum_{k=1}^{M} w_{ik}^* U(\theta; Y_{i,obs}, Y_{i,mis}^{*(k)}); \hat{\zeta}\right\} = 0,$$

where $\delta_i = \prod_{j=1}^{p} r_{ij}$. The fractional weights are constructed to satisfy

$$\sum_{k=1}^{M} w_{ik}^* U\left(\theta; Y_{i,obs}, Y_{i,mis}^{*(k)}\right) \cong E\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}\}$$

as closely as possible.

However, for multivariate data, it is not easy to find a joint distribution family $\{F_\zeta(Y) : \zeta \in \Omega\}$ correctly. If the joint distribution family $\{F_\zeta(Y) : \zeta \in \Omega\}$ is misspecified, the PFI can lead to

biased estimation and inference. All aforementioned concerns motivate us to consider a more robust fractional imputation method using Gaussian mixture model, which covers a wider class of parametric models.

5.3 Proposed Method

We assume that the random vector Y follows a Gaussian mixture model

$$f(Y;\alpha,\zeta)=\sum_{g=1}^{G}\alpha_g f(Y;\zeta_g), \tag{5.5}$$

where G is the number of mixture component, α_g is the mixture proportion satisfying $\sum_{g=1}^{G}\alpha_g=1$, $\zeta_g=\{\mu_g,\Sigma\}$ are the parameters belonging to the g-th Gaussian mixture distribution and $f(\cdot;\zeta_g)$ is the density function of multivariate normal distribution with parameter ζ_g. Here, we recommend using the same Σ across all components to get a parsimonious model. Note that, if the true model $F_0(Y)$ is one of the mixture components, then the mixture distribution should converge to the true distribution $F_0(Y)$ asymptotically.

To formulate the proposal, define the group indicator vector $Z=(z_1,z_2,\cdots,z_G)$, where $z_g=1$ and $z_j=0$ for all $j\neq g$, if sample unit belongs to the g-th group. Note that Z is a latent variable with parameter $\text{pr}(z_g=1)=\alpha_g$, satisfying $\sum_{g=1}^{G}\alpha_g=1$. Without loss of generality, we assume $\alpha_1\leq\alpha_2\leq\cdots\leq\alpha_G$ to avoid the non-identification issue. Now, we can express

$$f(Y)=\sum_{g=1}^{G}\text{pr}(z_g=1)f(Y\mid z_g=1),$$

which leads to the marginal distribution of Y in (5.5). To estimate ζ, the EM algorithm (Dempster et al., 1977) can be used under the complete observations of Y_i. If $\{(Z_i,Y_i)\}_{i=1}^{n}$ were fully observed, we could use the joint log-likelihood function

$$l_n(\alpha,\zeta)=\sum_{i=1}^{n}\sum_{g=1}^{G}z_{ig}\left\{\log\alpha_g+\log f(Y_i\mid z_{ig}=1;\zeta_g)\right\}. \tag{5.6}$$

Using (5.6), the EM algorithm of estimating α and ζ under the complete observations of $\{Y_1,\cdots,Y_n\}$ can be described as follows:

E-step: Compute the conditional expectation of the complete log-likelihood function in (5.6), given $\{Y_1, Y_2, \cdots, Y_n\}$ and the current estimators, $\alpha^{(t)}$ and $\zeta^{(t)}$, to obtain

$$Q(\alpha, \zeta \mid \alpha^{(t)}, \zeta^{(t)}) = E\left\{ l_n(\alpha, \zeta) \mid Y_1, \cdots, Y_n; \alpha^{(t)}, \zeta^{(t)} \right\}.$$

Since $l_n(\alpha, \zeta)$ is a linear function of z_{ig}, we can express

$$Q(\alpha, \zeta \mid \alpha^{(t)}, \zeta^{(t)}) = \sum_{i=1}^{n} \sum_{g=1}^{G} p_{ig}^{(t)} \left\{ \log \alpha_g + \log f(Y_i \mid z_{ig} = 1; \zeta_g) \right\}, \tag{5.7}$$

where

$$p_{ig}^{(t)} = \frac{f(Y_i \mid z_{ig} = 1; \zeta_g^{(t)}) \alpha_g^{(t)}}{\sum_{g=1}^{G} f(Y_i \mid z_{ig} = 1; \zeta_g^{(t)}) \alpha_g^{(t)}}$$

is the $t-$th estimate of $p_{ig} = \text{pr}(z_{ig} = 1 \mid Y_i)$.

M-step: Update the parameters by maximizing the conditional expectation of the complete log-likelihood function, in the sense of

$$(\alpha^{(t+1)}, \zeta^{(t+1)}) = \underset{\alpha, \zeta}{\text{argmax}}\, Q(\alpha, \zeta \mid \alpha^{(t)}, \zeta^{(t)}).$$

However, in addition to latent variable Z, Y is subject to missingness. Thus, to handle item nonresponse, we propose to use the fractional imputation method to impute the missing values. Note that, the joint predictive distribution of (Y_{mis}, Z) given Y_{obs} can be written as

$$f(Y_{mis}, Z \mid Y_{obs}) = f(Z \mid Y_{obs}) f(Y_{mis} \mid Y_{obs}, Z), \tag{5.8}$$

which implies that the prediction model for Y_{mis} is

$$f(Y_{mis} \mid Y_{obs}) = \sum_{g=1}^{G} \text{pr}(z_g = 1 \mid Y_{obs}) f(Y_{mis} \mid Y_{obs}, z_g = 1). \tag{5.9}$$

The first part in (5.9), which is $\text{pr}(z_g = 1 \mid Y_{obs})$, can be obtained by

$$\text{pr}(z_g = 1 \mid Y_{obs}) = \frac{f(Y_{obs} \mid z_g = 1) \alpha_g}{\sum_{g=1}^{G} f(Y_{obs} \mid z_g = 1) \alpha_g}$$

where $Y_{obs} \mid (z_g = 1)$ is normal. The second part $Y_{mis} \mid (Y_{obs}, z_g = 1)$ is also normal. Therefore, the proposed **f**ractional **i**mputation using **G**a**u**ssian mixtu**r**e mod**e**ls (FIGURE) can be described as

I-step: To generate $Y^*_{i,mis}$ from $f(Y_{i,mis} \mid Y_{i,obs}; \alpha^{(t)}, \zeta^{(t)})$ in (5.9), we use the following two-step method:

Step 1: Compute

$$p_{ig}^{(t)} = \frac{f(Y_{i,obs} \mid z_{ig} = 1; \zeta_g^{(t)})\alpha_g^{(t)}}{\sum_{g=1}^{G} f(Y_{i,obs} \mid z_{ig} = 1; \zeta_g^{(t)})\alpha_g^{(t)}},$$

where $f(Y_{i,obs} \mid z_{ig} = 1; \zeta_g)$ is the marginal density of $Y_{i,obs}$ derived from $(Y_{i,obs}, Y_{i,mis}) \mid (z_{ig} = 1) \sim N(\mu_g, \Sigma)$.

Step 2: Generate $Y^*_{i,mis}$ from

$$f(Y_{i,mis} \mid Y_{i,obs}; \alpha^{(t)}, \zeta^{(t)}) = \sum_{g=1}^{G} p_{ig}^{(t)} f(Y_{i,mis} \mid Y_{i,obs}, z_{ig} = 1; \zeta_g^{(t)}), \tag{5.10}$$

where $f(Y_{i,mis} \mid Y_{i,obs}, z_{ig} = 1; \zeta_g)$ is the conditional distribution derived from $(Y_{i,obs}, Y_{i,mis}) \mid z_{ig} = 1 \sim N(\mu_g, \Sigma)$. To generate M imputed values from (5.10), we first let $(M_1^{(t)}, M_2^{(t)}, \cdots, M_G^{(t)}) \sim$ Multinomial$(M; \underset{\sim i}{p}^{(t)})$, where $\underset{\sim i}{p}^{(t)} = (p_{i1}^{(t)}, \cdots, p_{iG}^{(t)})$. For each $g = 1, 2, \cdots, G$, we generate M_g independent realizations of $Y^*_{i,mis}$, say $\left\{Y^{*(g1)}_{i,mis}, Y^{*(g2)}_{i,mis}, \cdots, Y^{*(gM_g)}_{i,mis}\right\}$, from the conditional distribution $f(Y_{i,mis} \mid Y_{i,obs}, z_{ig} = 1)$, which is also normal.

W-step: Compute the fractional weights for $Y^{*(gj)}_{i,mis}$ as

$$w^*_{igj(t)} = p_{ig}^{(t)} \frac{1}{M_g^{(t)}}.$$

Using $(w^*_{igj(t)}, Y^{*(gj)}_{i,mis})$, we can compute

$$Q^*(\alpha, \zeta \mid \alpha^{(t)}, \zeta^{(t)}) = \sum_{i=1}^{n} \sum_{g=1}^{G} \sum_{j=1}^{M_g^{(t)}} w^*_{igj(t)} \left\{\log \alpha_g + \log f(Y_i^{*(gj)} \mid z_{ig} = 1; \zeta_g)\right\}, \tag{5.11}$$

where $Y_i^{*(gj)} = (Y_{i,obs}, Y^{*(gj)}_{i,mis})$. If $\delta_i = 1$, then $Y_i^{*(gj)} = Y_i$.

M-step: Update the parameters by maximizing (5.11) with respect to (α, ζ).

Repeat *I-step* and *M-step* until the convergence is achieved.

Then, the final estimator, say $\hat{\theta}_{FIGURE}$, of θ can be obtained by solving the fractionally imputed estimating equation, given by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{M_g} w_{igj}^{*} U(\theta; Y_i^{*(gj)}) = 0, \tag{5.12}$$

where w_{igj}^{*} are the final fractional weights and M_g are the final imputation sizes.

Remark 5.1. *We now briefly discuss variance estimation of $\hat{\theta}_{FIGURE}$. To estimate the variance of $\hat{\theta}_{FIGURE}$, the replication method can be used. First note that, the fractional weight assigned to $Y_i^{*(gj)}$ is*

$$w_{igj}^{*} = \hat{p}_{ig} M_g^{-1} := \hat{p}_{ig}\hat{\pi}_{2j|ig},$$

where $\hat{p}_{ig}$ is obtained from

$$\hat{p}_{ig} = \frac{f(Y_{i,obs} \mid z_{ig} = 1; \hat{\zeta}_g)\hat{\alpha}_g}{\sum_{g=1}^{G} f(Y_{i,obs} \mid z_{ig} = 1; \hat{\zeta}_g)\hat{\alpha}_g}. \tag{5.13}$$

Thus, the k-th replicate of w_{igj}^{*} can be obtained by

$$w_{igj}^{*(k)} = \hat{p}_{ig}^{(k)}\hat{\pi}_{2j|ig}^{(k)}, \tag{5.14}$$

where $\hat{p}_{ig}^{(k)}$ is obtained from (5.13) using $\hat{\zeta}^{(k)}$ and $\alpha_g^{(k)}$, the $k-$th replicate of $\hat{\zeta}$ and $\hat{\alpha}_g$ respectively, and

$$\hat{\pi}_{2j|ig}^{(k)} \propto \frac{f(Y_{i,mis}^{*(gj)} \mid Y_{i,obs}, z_{ig} = 1; \hat{\zeta}_g^{(k)})}{f(Y_{i,mis}^{*(gj)} \mid Y_{i,obs})}$$

and $\sum_{g=1}^{G}\hat{\pi}_{2j|ig}^{(k)} = 1$. The calculation of $\hat{\pi}_{2j|ig}^{(k)}$ is based on the idea of importance sampling. Construction of replicate fractional weights using importance sampling idea has been used in Berg et al. (2016).

The replicate parameter estimates $(\hat{\alpha}^{(k)}, \hat{\zeta}^{(k)})$ are computed by maximizing

$$l_{obs}^{(k)}(\alpha, \zeta) = \sum_{i=1}^{n} w_i^{(k)} \log f_{obs}(Y_{i,obs}; \alpha, \zeta) \tag{5.15}$$

respect to (α, ζ), where

$$f_{obs}(Y_{i,obs}; \alpha, \zeta) = \sum_{g=1}^{G} \alpha_g f(Y_{i,obs} \mid z_{ig} = 1; \zeta_g),$$

and $w_i^{(k)}$ *is the k-th replicate of* $w_i = n^{-1}$. *The maximizer of* $l_{obs}^{(k)}(\alpha, \zeta)$ *in (5.15) can be obtained by applying the same EM algorithm using replicate weights and replicate fractional weights in the M-step. There is no need to repeat I-step. Variance estimation for* $\hat{\theta}_{FIGURE}$ *can be obtained by computing the k-th replicate of* $\hat{\theta}_{FIGURE}$ *from*

$$\sum_{i=1}^{n} w_i^{(k)} \sum_{g=1}^{G} \sum_{j=1}^{M_g} w_{igj}^{*(k)} U(\theta; Y_i^{*(gj)}) = 0.$$

Remark 5.2. *In survey sampling, let* $\{(Y_1, w_1), (Y_2, w_2), \cdots, (Y_n, w_n)\}$ *represent the finite samples, where* w_i *are the sampling weights. The proposed FIGURE method can be applied to handle multivariate missingness under survey data. I-step is the same with IID setup. However, W-step is adapted to*

$$w_{igj(t)}^{*} = w_i p_{ig}^{(t)} \frac{1}{M_g}. \tag{5.16}$$

Note that $Q^*(\alpha, \zeta \mid \alpha^{(t)}, \zeta^{(t)})$ *in (5.11) is a pseudo log-likelihood function using (5.16). M-step is to maximize the pseudo log-likelihood function respect to* (α, ζ).

5.4 Asymptotic Theory

In our proposed FIGURE method, we assume G is fixed. If G is very large, the proposed mixture model may be subject to overfitting and increase its variance. If G is small, then the approximation of the true distribution cannot provide accurate prediction due to bias. Hence, we can allow G to depend on the sample size n, say $G = G(n)$. The choice of G under complete data has been well explored in the literature. The popular method are based on Bayesian information criterion (BIC) and Akaike's information criterion (AIC). See Wallace and Dowe (1999), Oliver et al. (1996), Windham and Cutler (1992), Schwarz et al. (1978), Fraley and Raftery (1998) and Dasgupta and Raftery (1998). The alternative way of using SCAD penalty (Fan and Li, 2001) is studies in Chen and Khalili (2008) and Huang et al. (2017). The resampling methods, such as cross-validation and bootstrap, can also be used to choose the number of mixture components. See McLachlan (1987) and Smyth (2000). Here, we propose to use the Bayesian information criterion to select G. Under

multivariate missingness, we do not have the complete log-likelihood function. Thus, we propose to use the observed log-likelihood function to serve the role of the complete log-likelihood function in the information criterion, in the sense that

$$\text{BIC} = -2\sum_{i=1}^{n}\log\left\{\sum_{g=1}^{G}\hat{\alpha}_g f(Y_{i,obs};\hat{\zeta}_g)\right\} + \{G-1+Gp+p(p+1)/2\}\log n, \tag{5.17}$$

under the assumption of $\Sigma_g = \Sigma$, where $(\hat{\alpha},\hat{\zeta})$ are the estimators obtained from the proposed method.

Considering the generalized penalties, we can rewrite (5.17) as

$$\text{BIC}(G) = -2\sum_{i=1}^{n}\log\left\{\sum_{g=1}^{G}\hat{\alpha}_g f(Y_{i,obs};\hat{\zeta}_g)\right\} + \log n\phi(G), \tag{5.18}$$

where $\phi(G)$ is a monotone increasing function of G. In (5.17), $\phi(G) = G+Gp$ if ignoring constants. In this section, we establish first the consistency of model selection using (5.18) under the Gaussian mixture model assumption.

Assume that samples $\{Y_1, Y_2, \cdots, Y_n\}$ are IID realizations from $f_0(Y) = \sum_{g=1}^{G^o}\alpha_g^o f(Y;\zeta_g^o)$, where (G^o,α^o,ζ^o) are true parameter values. For $\zeta_g^o = (\mu_g^o,\Sigma^o)$, we need the following regularity assumptions:

($A1$) The mean vectors for each mixture component is bounded uniformly, in the sense of $\|\mu_g^o\| \le C_1$, for $g = 1,2,\cdots,G^o$.

($A2$) $\|\Sigma^o\| \le C_2$. Furthermore, the smallest eigenvalue of Σ^o is positive.

The first assumption means the boundedness of the first moment. Assumption ($A2$) is to make sure that Σ^0 is bounded and nonsingular. Both assumptions are commonly used.

To establish the model consistency, we furthermore make the additional assumptions on the missingness mechanism:

($A3$) The response rate for y_j is bounded below from 0, say $\sum_{i=1}^{n} r_{ij}/n \ge C_3$, for $j = 1,2,\cdots,p$, where $C_3 >$.

($A4$) The response mechanism is MAR as defined in (5.3).

The following theorem shows that the true number of mixture components can be selected by minimizing BIC(G) in (5.18) consistently.

Theorem 5.1. *Assume the true density f_0 is the Gaussian mixture model, satisfying A1–A2. Let $\hat{G}$ be the minimizer of* BIC(G) *in (5.18). Under assumptions A3–A4, we have*

$$\Pr(\hat{G} = G^o) \to 1,$$

in probability, where G^o is the true number of mixture components.

The proof of Theorem 5.1 is shown in Appendix 5.9. For any continuous joint distribution, the selection of G using (5.18) can find the true GMM asymptotically.

Now, we establish the following lemma to measure how well GMM can approximate the arbitrary density function. We furthermore make additional assumptions about the true density function f_0. Use E_0 denote the expectation respect to f_0.

($A5$) Assume $f_0(Y)$ is continuous.

($A6$) Assume $E_0\{\partial f(Y)/\partial\alpha\} < \infty$ and $E_0\{\partial f(Y)/\partial\mu\} < \infty$, where $f(Y) = \sum_{g=1}^{G} \alpha_g f(Y; \mu_g, \Sigma)$. Moreover, assume $E_0\{f(Y)^{-2}\} < \infty$.

($A7$) $\int Y^2 f_0(Y) < \infty$.

Lemma 5.1. *Under assumptions (A5)–(A7) and MAR, for any $\epsilon > 0$, there exist $G = \epsilon^{-\gamma}$, such that*

$$\|f_0 - \hat{f}\|_1 = O(\epsilon), \tag{5.19}$$

$$\text{var}(f_0 - \hat{f}) = O(\epsilon^{-\gamma} n^{-1}), \tag{5.20}$$

with probability one, where $\hat{f}(Y) = \sum_{g=1}^{G} \hat{\alpha}_g f(Y; \hat{\mu}_g, \hat{\Sigma})$ is obtained from the proposed method in §5.3, $\gamma > 0$ depends on f_0 and $\|f_0 - \hat{f}\|_1 = \int |f_0(Y) - \hat{f}(Y)| f_0(Y) \mathrm{d}Y$.

The proof of Lemma (5.1) is presented in Appendix 5.11. If f_0 is a density function of the Gaussian mixture model, then $\gamma = 0$. Then, our proposed BIC(G) in Theorem 5.1 can select the

true model consistently. For any f_0 satisfies $(A5)$–$(A7)$ and is not a finite Gaussian mixture model, the bias can goes to 0 as $G \to \infty$ from (5.19). The variance will increase as $G \to \infty$ from (5.20) for fixed n. There is a trade-off between bias and variance for the divergence case ($\gamma > 0, G \to \infty$).

Using Lemma 5.1, we further establish the $\sqrt{n}$-consistency of $\hat{\theta}_{FIGURE}$. The following assumptions are the sufficient conditions to obtain the $\sqrt{n}$-consistency.

$(A8)$ $E_0\left\{U^2(\theta; Y_i)\right\} < \infty$.

$(A9)$ $\gamma \in (0, 2)$.

$(A10)$ $\epsilon = O(n^{-1/(2-\Delta)})$, for any $\Delta \in (0, 2)$.

Theorem 5.2. *Under assumptions (A5)–(A10), $\gamma + \Delta < 2$ and MAR, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{M_g} w_{igj}^{*} U(\theta; Y_i^{*(gj)}) \cong J_1 + o_p(n^{-1/2}), \tag{5.21}$$

where

$$J_1 = \frac{1}{n}\sum_{i=1}^{n} E_0\left\{U(\theta; Y_i) \mid Y_{i,obs}\right\},$$

if $M = \min_g\{M_g\} \to \infty$. Furthermore, we have

$$\sqrt{n}(\hat{\theta}_{FIGURE} - \theta_0) \to N(0, \Sigma), \tag{5.22}$$

for some Σ which is positive definite and θ_0 satisfies $E_0\left\{U(\theta_0; Y)\right\} = 0$.

The proof of (5.21) is shown in Appendix 5.12 and (5.22) is the following result from (5.21). From Theorem 5.2, we have $G = O(n^{\gamma/(2-\Delta)}) \to \infty$ with the rate smaller than n. Thus, under divergence case, our proposed method still enjoys $\sqrt{n}$-consistency.

5.5 Extension

In Section 5.3, we assume that Y is fully continuous. However, in practice, many categorical data, such as demographic variables, can be used to build an imputation model. To extend

the proposed FIGURE method to incorporate the categorical variables, we propose the following conditional FIGURE (CFIGURE) method.

To introduce the CFIGURE method, we first introduce the conditional GMM. Suppose that (X, Y) is a random vector where X is discrete and Y is continuous. To obtain the conditional GMM, we assume that Z satisfies

$$f(Y \mid X, Z) = f(Y \mid Z), \tag{5.23}$$

in the sense that Z is a partition of the sample such that Y is homogeneous within each group. Furthermore, we assume that $f(Y \mid z_g = 1)$ follows a Gaussian distribution. Combining these assumptions, we have the following conditional GMM

$$\begin{aligned} f(Y \mid X) &= \sum_{g=1}^{G} \mathrm{pr}(z_g = 1 \mid X) f(Y \mid X, z_g = 1) \\ &= \sum_{g=1}^{G} \mathrm{pr}(z_g = 1 \mid X) f(Y \mid z_g = 1) \\ &= \sum_{g=1}^{G} \alpha_g(X) f(Y \mid z_g = 1), \end{aligned} \tag{5.24}$$

where $\alpha_g(X)$ is the conditional probability $\mathrm{pr}(z_g = 1 \mid X)$ and $f(Y \mid z_g = 1)$ is the density function of the normal distribution with parameter $\zeta_g = \{\mu_g, \Sigma\}$.

To make group indicator vector Z based on the fully observed samples $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, the following fractional imputation method can be applied. Similarly to §5.3, if $Z_1, \cdots, Z_n$ were observed, then the complete log-likelihood function could be written as

$$l_n(\alpha, \zeta) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left\{ \log \alpha_g(X_i) + \log f(Y_i \mid z_{ig} = 1; \zeta_g) \right\}.$$

Moreover, since X_i and Z_i are discrete, $\alpha_g(X_i) = \mathrm{pr}(z_{ig} = 1 \mid X_i)$ can be estimated directly from the empirical distribution. However, Z is latent. The predictive model of Z can be obtained by

$$\mathrm{pr}(z_g = 1 \mid Y, X) = \frac{f(Y \mid z_g = 1)\mathrm{pr}(z_g = 1 \mid X)}{\sum_{g=1}^{G} f(Y \mid z_g = 1)\mathrm{pr}(z_g = 1 \mid X)}, \tag{5.25}$$

due to (5.23). The parameter estimation for the conditional GMM (CGMM) can be described as follows:

E-step: Given the current values of parameters, $\alpha_g^{(t)}(X)$ and $\zeta_g^{(t)}$, using (5.25), compute the predictive probabilities as

$$p_{ig}^{(t)} \propto \alpha_g^{(t)}(X_i) f(Y_i \mid z_{ig} = 1; \zeta_g^{(t)}), \tag{5.26}$$

where $\sum_{g=1}^{G} p_{ig}^{(t)} = 1$. Then, we can compute the conditional expectation of $l_n(\alpha, \zeta)$ as

$$Q(\alpha, \zeta \mid \alpha^{(t)}, \zeta^{(t)}) = \sum_{i=1}^{n} \sum_{g=1}^{G} p_{ig}^{(t)} \left\{ \log \alpha_g(X_i) + \log f(Y_i \mid z_{ig} = 1; \zeta_g) \right\}. \tag{5.27}$$

M-step: Update the proportion vector by

$$\alpha_g^{(t+1)}(X_i) = \frac{\sum_{\{j: X_j = X_i\}} p_{jg}^{(t)}}{\sum_{g=1}^{G} \sum_{\{j: X_j = X_i\}} p_{jg}^{(t)}}. \tag{5.28}$$

The parameters ζ can be updated by maximizing $Q(\alpha^{(t+1)}, \zeta \mid \alpha^{(t)}, \zeta^{(t)})$ in (5.27) respect to ζ.

Next, we extend the above EM algorithm under CGMM to incorporate item nonresponse. For simplicity, we only consider that X is fully observed and Y is subject to missingness. Under (5.23), the predictive model of $Y_{i,mis}$ can be expressed as

$$f(Y_{i,mis} \mid Y_{i,obs}, X_i) = \sum_{g=1}^{G} \text{pr}(z_{ig} = 1 \mid Y_{i,obs}, X_i) f(Y_{i,mis} \mid Y_{i,obs}, z_{ig} = 1), \tag{5.29}$$

where $f(Y_{i,mis} \mid Y_{i,obs}, z_{ig} = 1)$ can be derived from $(Y_{i,obs}, Y_{i,mis}) \mid (z_{ig} = 1) \sim N(\mu_g, \Sigma)$. Similarly to (5.25), the posterior probability of $z_{ig} = 1$ given observed data can be obtained as

$$\text{pr}(z_{ig} = 1 \mid Y_{i,obs}, X_i) = \frac{f(Y_{i,obs} \mid z_{ig} = 1) \text{pr}(z_{ig} = 1 \mid X_i)}{\sum_{g=1}^{G} f(Y_{i,obs} \mid z_{ig} = 1) \text{pr}(z_{ig} = 1 \mid X_i)}. \tag{5.30}$$

Therefore, the proposed CFIGURE can be summarized as follows:

I-step: Creating M imputed values of $Y_{i,mis}$ from (5.29) can be described as the following two steps.

Step 1: For each $g = 1, 2, \cdots, G$, given the current parameter values $(\alpha_g^{(t)}, \zeta_g^{(t)})$, the posterior probabilities of $z_{ig} = 1$ given $(Y_{i,obs}, X_i)$ can be obtained from

$$p_{ig}^{(t)} = \frac{f(Y_{i,obs} \mid z_{ig} = 1; \zeta_g^{(t)}) \alpha_g^{(t)}(X_i)}{\sum_{g=1}^{G} f(Y_{i,obs} \mid z_{ig} = 1; \zeta_g^{(t)}) \alpha_g^{(t)}(X_i)}.$$

Step 2: Generate M imputed values of $Y_{i,mis}$ following the same procedures in *I-step* in §5.3.

W-step: Update the fractional weights for $Y_i^{*(gj)} = (Y_{i,obs}, Y_{i,mis}^{*(gj)})$ as

$$w_{igj(t)}^{*} = \frac{p_{ig}^{(t)}}{M_g},$$

for $j = 1, 2, \cdots M_g$ and $\sum_{g=1}^{G} M_g = M$.

M-step: Update the parameter values by maximizing

$$Q^{*}(\alpha, \zeta \mid \alpha^{(t)}, \zeta^{(t)}) = \sum_{i=1}^{n} \sum_{g=1}^{G} \sum_{j=1}^{M_g} w_{igj(t)}^{*} \left\{ \log \alpha_g(X_i) + \log f(Y_i^{*(gj)} \mid z_{ig} = 1; \zeta_g) \right\},$$

respect to (α, ζ).

Repeat *I-step* to *M-step* iteratively until convergence is obtained. The final estimator of θ can be obtained by solving the fractionally imputed estimating equation in (5.12). Note that the proposed CFIGURE method builds the proportion vector of mixture components into a function of auxiliary variable and assumes that mixture components share the same mean and variance structure. Thus, the proposed method is useful in borrowing information across X. Moreover, the auxiliary information is incorporated to build a more flexible class of joint distributions.

5.6 Numerical Studies

In this section, we conducted two numerical studies to evaluate the performance of the proposed method. The first simulation study is used to check the performance of FIGURE under multivariate continuous variables. Heavy tailed, skewed and nonlinear distributions are used to demonstrate the efficiency and robustness of the proposed method. The second simulation study considers the case of multivariate mixed categorical and continuous variables.

5.6.1 Simulation Study I

The design for the first simulation study can be described as a 4×2 factorial design, where the two factors are outcome model and response mechanism. We consider the following models.

M1: $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$ follows a mixture distribution with density $f(Y) = \sum_{g=1}^{3} \alpha_g f_g(Y)$, where $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.3, 0.4)$ and $f_g(Y)$ is a density function for multivariate normal distribution with mean μ_g and variance Σ. Let $\mu_1 = (-3, -3, -3, -3), \mu_2 = (1, 1, 1, 1), \mu_3 = (5, 5, 5, 5)$ and

$$\Sigma = \begin{pmatrix} 1 & 0.8 & 0.8^2 & 0.8^3 \\ 0.8 & 1 & 0.8 & 0.8^2 \\ 0.8^2 & 0.8 & 1 & 0.8 \\ 0.8^3 & 0.8^2 & 0.8 & 1 \end{pmatrix}. \tag{5.31}$$

M2: Use the same model as *M1* except for $f_1(Y)$, where $f_1(Y)$ is the density for t distribution with degree freedom 5 and non-centrality -3.

M3: Let $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$, where $x_{ij}, j = 1, 2, 3, 4$, are independently generated from Gamma$(1, 1)$. Let $Y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})$, where $y_{i1} = x_{i1}, y_{i2} = x_{i1} + x_{i2}, y_{i3} = x_{i1} + x_{i2} + x_{i3}$ and $y_{i4} = x_{i1} + x_{i4}$.

M4: Generate $x_i \sim N(1, 1)$ independently. Let $Y_i = (x_i, x_i^2, x_i^3, x_i^4)$.

Under *M1*, the proposed FIGURE method correctly specifies the joint model. Non-centered t distribution is used in *M2* to check the robustness of the proposed method to the outliers and heavy tails. *M3* and *M4* are used to check the performance of FIGURE under skewness and nonlinearity, respectively.

The sample size for each realized sample is $n = 500$. Once the complete sample is obtained, we apply the following two response mechanisms to create two separate incomplete samples.

1. *MCAR*: For $Y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})$, assume y_{i1} are fully observed. For $j > 1$, we use simple random sampling to select 20% to make missingness equally for each item. There are about 50% complete data overall. The response mechanism is missing completely at random (MCAR).

2. *MAR*: Define

$$\pi_i = \frac{\exp(-0.5 + 0.5y_{i1})}{1 + \exp(-0.5 + 0.5y_{i1})}.$$

For $y_{ij}, j = 2, 3, 4$, we select 20% of the sample independently to make missingness with the selection probabilities equal to π_i. Since we assume y_{i1} are fully observed, the response mechanism is MAR.

For each realized incomplete samples, we apply the following methods:

[*Full*]: As a benchmark, we use the full samples to estimate parameters. 95% confident intervals are constructed using full sample standard errors.

[*CC*]: Only use the complete cases to estimate parameters and construct confidence intervals.

[*MICE*]: Apply multivariate imputation by chained equations (Buuren and Groothuis-Oudshoorn, 2011). The variance estimators are obtained using Rubin's formula in Rubin (2004) and confidence intervals are built using the asymptotic normality.

[*FIGURE*]: The proposed method where the number of components G is selected using the BIC in (5.17). The inference is implemented using the variance estimator presented in Remark 5.1.

The parameters of interest are sample means and sample proportions. For $Y = (y_1, y_2, y_3, y_4)$, define $\theta_2 = E(y_2), \theta_3 = E(y_3)$ and $\theta_4 = E(y_4)$. For outcome model *M1*, define $P_2 = \Pr(y_2 < -2), P_3 = \Pr(y_3 < -2), P_4 = \Pr(y_4 < -2)$ and $P_2 = \Pr(y_2 < -3), P_3 = \Pr(y_3 < -3), P_4 = \Pr(y_4 < -3)$ in *M2*. For *M3*, define $P_2 = \Pr(y_2 < 2), P_3 = \Pr(y_3 < 3), P_4 = \Pr(y_4 < 2)$ and $P_2 = \Pr(y_2 < 0.6), P_3 = \Pr(y_3 < 1.5), P_4 = \Pr(y_4 < 1)$ in *M4*. The simulation is repeated for $B = 2,000$ times.

To evaluate the above methods, the relative mean square error (RMSE) is defined as

$$\text{RMSE} = \frac{\text{MSE}_{\text{method}}}{\text{MSE}_{\text{Full}}} \times 100, \tag{5.32}$$

where $\text{MSE}_{\text{method}}$ is the mean square error of the parameters of applying method and MSE_{Full} is the mean square error of the parameters of using full samples. The simulation results of RMSE are

presented in Table 5.1. The average of coverage probabilities and interval length of $2,000$ Monte Carlo 95% confidence intervals are presented in Table 5.4 in Appendix 5.10.

From Table 5.1, when the outcome model is the Gaussian mixture model (*M1*), all methods the are consistent under MCAR. MICE and FIGURE have almost the same performance in term of relative mean square errors (RMSEs). However, CC is less efficient due to smaller sample size with ignoring the missingness under MCAR. When the response mechanism is MAR, the CC method is biased, which leads to large RMSE. When the outcome model has heavy tails and outliers (*M2*), FIGURE is slightly better than MICE under both MCAR and MAR response mechanisms. When the outcome model is skewed (*M3*), FIGURE has almost the same RMSE with MICE for mean estimators, but outperforms MICE for proportion estimation. Thus, our proposed FIGURE method preserves the correlation structure better than MICE. When the outcome model has nonlinear mean curves (*M4*), FIGURE has much smaller RMSE than MICE for proportion estimators. Thus, the proposed FIGURE is more robust and efficient for general purpose estimation.

Interestingly, imputed estimators are sometimes more efficient than the full sample estimators. This phenomenon, called superefficiency (Meng, 1994), can happenïijŇ when the method-of-moment is used in the full sample estimator. Yang and Kim (2016a) give a rigorous theoretical justification for this phenomenon.

From the coverage probabilities in Table 5.4, the proposed replicate inference procedure in Remark 5.1 estimates confidence intervals consistently. Moreover, for *M4*, both FIGURE and MICE suffer under-coverage for proportion estimation. However, FIGURE provides wider confidence intervals and better coverage probabilities than MICE for proportion estimators in all cases. Therefore, FIGURE is more robust to model misspecification than MICE.

5.6.2 Simulation Study II

The second simulation study is used to check the performance of the proposed CFIGURE in §5.5 under mixed type data. The outcome model can be generated as follows:

M5: $V_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$ are independently generated from Gaussian mixture model with density function $f(V) = \sum_{g=1}^{G} \alpha_g f_g(V)$. Let the mixture proportion vector and mean vectors be the same with *M1*. However, we use

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.7^2 & 0.7^3 \\ 0.5 & 1 & 0.7 & 0.7^2 \\ 0.7^2 & 0.7 & 1 & 0.7 \\ 0.7^3 & 0.7^2 & 0.7 & 1 \end{pmatrix}$$

to reduce the correlation. Generate the auxiliary variables from

$$X_i = \begin{cases} 1 & \text{if } v_{i1} < 0 \\ 2 & \text{if } 0 \leq v_{i1} < 3 \\ 3 & \text{otherwise} \end{cases} \tag{5.33}$$

and $Y_i = (v_{i2}, v_{i3}, v_{i4})$.

M6: Generate the model indicators X_i independently using simple random sampling from $\{1, 2, 3\}$ with probabilities $(0.3, 0.3, 0.4)$. Let $Y_i = (y_{i1}, y_{i2}, y_{i3})$ follows a multivariate normal distribution. If $X_i = 1$, generate Y_i using mean $(-3, -3, -3)$ and variance Σ, where

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.49 \\ 0.5 & 1 & 0.7 \\ 0.49 & 0.7 & 1 \end{pmatrix}.$$

If $X_i = 2$, generate Y_i using mean $(1, 1, 1)$ and variance Σ. For all other X_i, use mean $(5, 5, 5)$ and Σ.

M7: Generate X_i using simple random sampling from $\{1, 2\}$ with probabilities $(0.7, 0.3)$ independently. If $X_i = 1$, then, Y_i is generated from multivariate normal distribution with mean $(1, 1, 1)$ and variance

$$\Sigma = \begin{pmatrix} 1 & 0.7 & 0.49 \\ 0.7 & 1 & 0.7 \\ 0.49 & 0.7 & 1 \end{pmatrix}.$$

For $X_i = 2$, let $Y_i = (y_{i1}, y_{i2}, y_{i3})$. Then, generate y_{ij} from Gamma distribution with a shape parameter 1 and a scale parameter 1 independently for $j \geq 1$.

M5 is simulated from GMM but using discretized y_{i1} as the auxiliary variables. In *M6*, X_i are the indicators of groups, which often happen in demographical variables. *M6* is mixture of Gaussian and Gamma distributions. We use *M7* to test the robustness of Gaussian assumption.

Suppose that X_i are fully observed and Y_i are subject to multivariate missingness. For the response mechanism, MCAR and MAR are applied. MCAR is the same as the simulation study I. For MAR, define

$$\pi_i = \frac{\exp(0.5 + 0.5X_i)}{1 + \exp(0.5 + 0.5X_i)}.$$

Then, we select 20% of items to make missingness using probabilities π_i for each $y_{ij}, j = 2, 3, 4,$. The overall response rate is approximately 50%.

The parameters of interest are mean and proportion estimators of Y . For all three models, let $\theta_1 = E(y_1), \theta_2 = E(y_2), \theta_3 = E(y_3)$. For *M5* and *M6*, let $P_1 = \Pr(y_1 < 0), P_2 = \Pr(y_2 < 0), P_3 = \Pr(y_3 < 0)$. For *M7*, let $P_1 = \Pr(y_1 < 1.5), P_2 = \Pr(y_2 < 1.5), P_3 = \Pr(y_3 < 1.5)$.

For comparison, we also apply MICE in Simulation Study I to each realized incomplete sample. We repeat simulation process $B = 2,000$ times. The simulation results are shown in Table 5.2 and Table 5.5.

Table 5.2 presents RMSE of MICE and CFIGURE. For the proposed CFIGURE, the key assumption is $\Pr(Y \mid X, z_g = 1) = \Pr(Y \mid z_g = 1)$. Under *M5*, the assumption of $\Pr(Y \mid X, z_g = 1) = \Pr(Y \mid z_g = 1)$ is violated. Even though the assumption is violated, the proposed CFIGURE has better performance of estimating proportions than MICE and similar RMSE for mean estimators under both MCAR and MAR. Under *M6*, the assumption of $\Pr(Y \mid X, z_g = 1) = \Pr(Y \mid z_g = 1)$ holds. The proposed CFGMM outperforms MICE to estimate proportions. MICE works well under *M5* and *M6*, since the normality holds and regression structure, which depends on Σ are constant across groups. Under *M7*, the Gaussian mixture assumption is violated and the mixture of independent gamma distributions destroy the regression structures. Thus, the proposed CFIGURE uniformly performs better than MICE.

Table 5.5 shows the average lengths of 95% confidence intervals and coverage probabilities using Jackknife method introduced in Remark 5.1. Table 5.5 demonstrates that the proposed Jackknife method can provides valid inferences.

5.7 Application

In this section, we apply the proposed method in §5.3 to a synthetic data that mimics monthly retail trade survey data at U.S. Census Bureau. The monthly retail trade survey data can be found in http://www.portal-stat.admin.ch/ices5/imputation-contest/. The sampling scheme is a stratified simple random sample without replacement sample with six strata: one certain (take-all) and five non-certainty strata. The sample sizes are computed using Neyman allocation.

The overview of the monthly retail trade survey data is presented in Figure 5.1. The overall

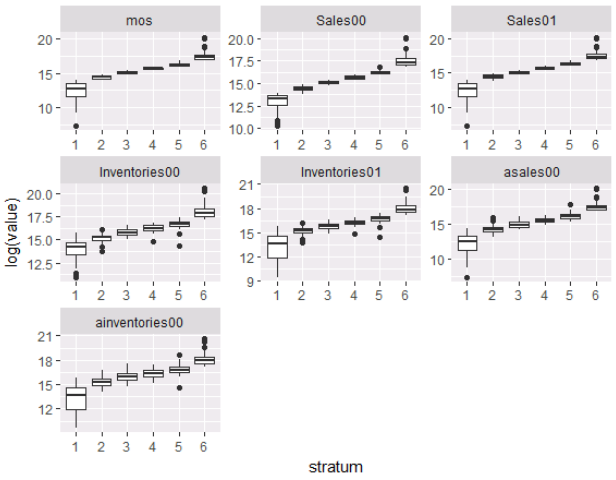


Figure 5.1: "mos" is frame measure of size; "Sales00" denotes current month sales for unit (subject to missing); "Asales00" is current month administrative data value for sales; "Sales01" means prior month sales for unit; "Inventories00" is current month inventories for unit (Subject to missing); "Ainventories00" is current month administrative data value for inventories; "Inventories01" is prior month inventories for unit.

response rate is approximately 71%. Current month sales and inventories are subject to missingness. We can find that this monthly retail trade data are highly skewed. The normal quantile-quantile plots are shown in Figure 5.2. From Figure 5.2, Gaussian assumption is violated and there exist

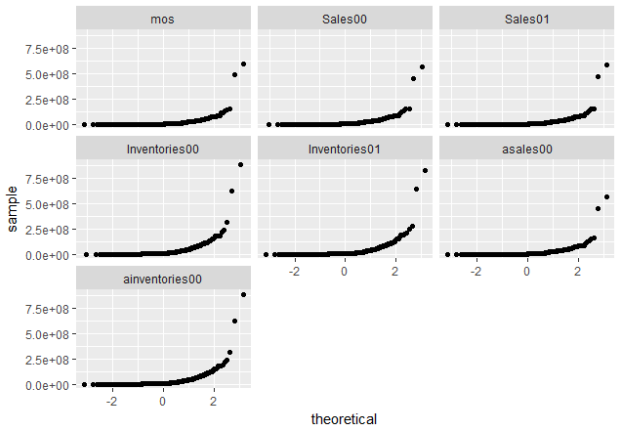


Figure 5.2: Quantile-quantile plots for the monthly retail trade survey data.

three extreme outliers.

To impute current month sales and inventories, we apply the proposed FIGURE method and MICE. After implementation, MICE fails to converge due to high correlations. See the correlation plot in Figure 5.3. Therefore, we only present the final results using the proposed method. The

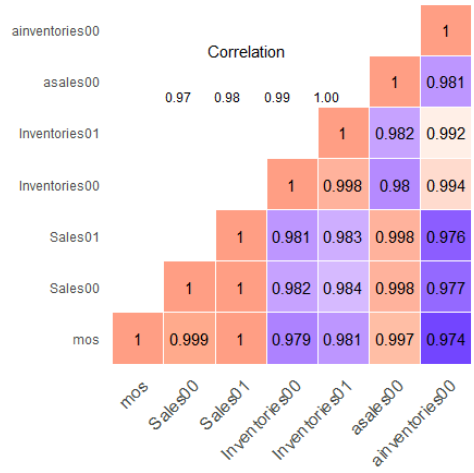


Figure 5.3: Correlation plot of the monthly retail trade survey data only using complete cases.

final results are shown in Table 5.3.

Comparing with the true population statistics, provided by U.S. Census Bureau, we can see that our proposed FIGURE method works well to preserve the correlation structure and handle

skewness and outliers. The 95% confidence intervals are also presented using Jackknife method in Remark 5.1. We can see that all 95% confidence intervals contain their true values.

5.8 Discussion

In this paper, we propose a semiparametric fractional imputation method using GMM to handle arbitrary multivariate missing data. The proposed method automatically selects mixture components and provides a unified framework for robust imputation. Even if the group size G can increase with sample size n, the resulting estimator is $\sqrt{n}$-consistency. We also extend the proposed method to incorporate categorical auxiliary variable. The flexible model assumption and efficient computation are main advantages of our proposed method. The proposed method is directly applicable in survey sample data. An R software package for the proposed method is under development.

5.9 Appendix A: Proof of Theorem 5.1

The outline of the proof can be described as the following two steps:

Step 1: Show $\Pr(\hat{G} > G^o) \to 0$ in probability.

Step 2: Show $\Pr(\hat{G} < G^o) \to 0$ in probability.

Thus, combining *Step 1* and *Step 2*, we can complete the proof.

Before we show the proof, let us define some notations first. From assumption $(A2)$, Σ^o is bounded and invertible. Thus, we can define $Y_i \leftarrow \{\Sigma^o\}^{-1/2} Y_i$. Therefore, without loss of generality, we consider the standardized samples $\{Y_i\}$. Then, $\zeta_g = \mu_g$.

Given $G = \hat{G}$, we can obtain $\{(\hat{\mu}_g, \hat{\alpha}_g)\}_{g=1}^{\hat{G}}$ from the proposed FIGURE. Similarly, if $G = G^o$, $\{(\hat{\mu}_g^o, \hat{\alpha}_g^o)\}_{g=1}^{\hat{G}^o}$ can be obtained.

Using Theorem 1 in Kim (2011), we can obtain the following lemma.

Lemma 5.2. *Under the regularity conditions in Kim (2011), given $G \geq G^o$, $\{(\hat{\mu}_g^o, \hat{\alpha}_g^o)\}_{g=1}^{\hat{G}^o}$ converge to true values with rate of $O_p(1/\sqrt{n})$.*

We first show *Step 1.* If $\hat{G} > G^o$, we assume the first G^o components are non-negligible. Thus, using Lemma 5.2, we have $\hat{\mu}_g \to \hat{\mu}_g^o$ and $\hat{\alpha}_g \to \alpha_g^o$ in probability, for $g = 1, 2, \cdots, G^o$. Moreover, $\hat{\alpha}_g \to 0$ in probability, for $g = G^o + 1, \cdots, \hat{G}$.

Using Taylor linearization, we have

$$\begin{aligned}
&-2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{\hat{G}} \hat{\alpha}_g f(Y_{i,obs}; \hat{\zeta}_g)\right\} \\
&= -2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{G^o} \alpha_g^o f(Y_{i,obs}; \zeta_g^o)\right\} + D_1^T \beta_1 + O_p(1),
\end{aligned}$$

where $\beta_1 = (\hat{\mu}_1 - \mu_1, \cdots, \hat{\mu}_{G^o} - \mu_{G^o}, \hat{\alpha}_1 - \alpha_1, \cdots, \hat{\alpha}_{G^o} - \alpha_{G^o}, \hat{\mu}_{G^o+1}, \cdots, \hat{\mu}_{\hat{G}}, \hat{\alpha}_{G^o+1}, \cdots, \hat{\alpha}_{\hat{G}})$ and

$$D_1 = \left.\frac{\partial\left[-2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs}; \zeta_g)\right\}\right]}{\partial(\alpha, \mu)}\right|_{\alpha=\alpha^o, \mu=\mu^o}.$$

Similarly, we can obtain that

$$\begin{aligned}
&-2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{G^o} \hat{\alpha}_g^o f(Y_{i,obs}; \hat{\zeta}_g^o)\right\} \\
&= -2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{G^o} \alpha_g^o f(Y_{i,obs}; \zeta_g^o)\right\} + D_2^T \beta_2 + O_p(1),
\end{aligned}$$

where $\beta_2 = (\hat{\mu}_1 - \mu_1, \cdots, \hat{\mu}_{G^o} - \mu_{G^o}, \hat{\alpha}_1 - \alpha_1, \cdots, \hat{\alpha}_{G^o} - \alpha_{G^o})$ and

$$D_2 = \left.\frac{\partial\left[-2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{G^o} \alpha_g^o f(Y_{i,obs}; \zeta_g^o)\right\}\right]}{\partial(\alpha, \mu)}\right|_{\alpha=\alpha^o, \mu=\mu^o}.$$

Note that, given true parameter values, the first $2G^o$ entries of D_1 are equal to the first $2G^o$ entries of D_2. Therefore, we have

$$\begin{aligned}
&-2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{\hat{G}} \hat{\alpha}_g f(Y_{i,obs}; \hat{\zeta}_g)\right\} + \log n\phi(\hat{G}) \\
&+2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{G^o} \hat{\alpha}_g^o f(Y_{i,obs}; \hat{\zeta}_g^o)\right\} - \log n\phi(G^o) \\
&= \log n\left\{\phi(\hat{G}) - \phi(G^o)\right\} + \left\{D_1^{[(2G^o+1):(2\hat{G})]}\right\}^T (\hat{\mu}_{G^o+1}, \cdots, \hat{\mu}_{\hat{G}}, \hat{\alpha}_{G^o+1}, \cdots, \hat{\alpha}_{\hat{G}})^T + O_p(1),
\end{aligned}$$

where $D_1^{[(2G^o+1):(2\hat{G})]}$ is the vector of $D1$ from $2G^o + 1$ to $2\hat{G}$.

We can show that

$$\frac{\partial \log\left\{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs};\zeta_g)\right\}}{\partial \alpha_g} = \frac{f(Y_{i,obs};\zeta_g)}{\log\left\{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs};\zeta_g)\right\}},$$
$$\frac{\partial \log\left\{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs};\zeta_g)\right\}}{\partial \mu_g} = \frac{\alpha_g}{\log\left\{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs};\zeta_g)\right\}} \frac{\partial f(Y_{i,obs};\zeta_g)}{\partial \mu_g}.$$

For $g > G^o$, $\alpha_g = 0$. Thus,

$$\frac{\partial \log\left\{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs};\zeta_g)\right\}}{\partial \mu_g} = 0.$$

Since the true model does not have g-th mixture component for $g > G^o$, we can let $\mu_g \to \infty$. Thus,

$$\frac{\partial \log\left\{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs};\zeta_g)\right\}}{\partial \alpha_g} \to 0.$$

Therefore, we can show that

$$\left\{D_1^{[(2G^o+1):(2\hat{G})]}\right\}^T (\hat{\mu}_{G^o+1},\cdots,\hat{\mu}_{\hat{G}},\hat{\alpha}_{G^o+1},\cdots,\hat{\alpha}_{\hat{G}})^T = o_p(1).$$

Overall, we show

$$\begin{aligned}
&-2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{\hat{G}} \hat{\alpha}_g f(Y_{i,obs};\hat{\zeta}_g)\right\} + \log n\phi(\hat{G}) \\
&+2\sum_{i=1}^{n} \log\left\{\sum_{g=1}^{G^o} \hat{\alpha}_g^o f(Y_{i,obs};\hat{\zeta}_g^o)\right\} - \log n\phi(G^o) \\
&= \log n\left\{\phi(\hat{G}) - \phi(G^o)\right\} + O_p(1).
\end{aligned}$$

Since $\phi(G)$ is a monotone increasing function, we have $\log n\left\{\phi(\hat{G}) - \phi(G^o)\right\} + O_p(1) > 0$ in probability as $n \to \infty$. Thus, G^o is the minimizer of (5.18), instead of $\hat{G}$, which completes the proof of *Step 1.*

Next step is to prove $\Pr(\hat{G} < G^o) \to 0$ in probability. To show that, we first introduce KullbackâĂŞLeibler (KL) divergence. For distributions F and Q of a continuous random variable, the KL divergence is defined as

$$\mathrm{KL}(F|Q) = \int f(x) \log \frac{f(x)}{q(x)} \mathrm{d}x.$$

The KL divergence is always non-negative. Applying KL divergence to the two Gaussian mixture density functions, we have

$$\int \sum_{g=1}^{G^o} \alpha_g^o f(Y_i;\zeta_g^o) \log \frac{\sum_{g=1}^{G^o} \alpha_g^o f(Y_i;\zeta_g^o)}{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_i;\zeta_g)} \mathrm{d}(Y_{i,obs}) \geq 0,$$

under complete data.

Under MAR assumption in $(A4)$ and non-empty observations in $(A3)$, we can show that

$$\int \sum_{g=1}^{G^o} \alpha_g^o f(Y_{i,obs};\zeta_g^o) \log \frac{\sum_{g=1}^{G^o} \alpha_g^o f(Y_{i,obs};\zeta_g^o)}{\sum_{g=1}^{\hat{G}} \alpha_g f(Y_{i,obs};\zeta_g)} \mathrm{d}(Y_{i,obs}) \geq 0, \tag{5.34}$$

Since $\hat{G} < G^o$, (5.34) is positive and denote it as $C_4(R_i) > 0$.

From Assumption $(A3)$, we can define $C_4 = \min_{R_i, \sum_j r_{ij}>0} C_4(R_i)$. Then, we can show that

$$\begin{aligned}
&-2\sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{\hat{G}} \hat{\alpha}_g f(Y_{i,obs};\hat{\zeta}_g) \right\} + \log n\phi(\hat{G}) + 2\sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{G^o} \hat{\alpha}_g^o f(Y_{i,obs};\hat{\zeta}_g^o) \right\} - \log n\phi(G^o) \\
&= 2\sum_{i=1}^{n} \log \frac{\sum_{g=1}^{G^o} \hat{\alpha}_g^o f(Y_{i,obs};\hat{\zeta}_g^o)}{\sum_{g=1}^{\hat{G}} \hat{\alpha}_g f(Y_{i,obs};\hat{\zeta}_g)} - \log n \left\{ \phi(G^o) - \phi(\hat{G}) \right\} \\
&\to 2\sum_{R_i} n(R_i) C_4(R_i) - \log n \left\{ \phi(G^o) - \phi(\hat{G}) \right\},
\end{aligned}$$

where $n(R_i)$ is the count of missing pattern R_i and $\sum_{R_i} n(R_i) = n$. Since

$$\sum_{R_i} n(R_i) C_4(R_i) \geq nC_4.$$

Since $nC_4 - \log n \left\{ \phi(G^o) - \phi(\hat{G}) \right\} > 0$ when n is large enough, we can conclude that

$$-2\sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{\hat{G}} \hat{\alpha}_g f(Y_{i,obs};\hat{\zeta}_g) \right\} + \log n\phi(\hat{G}) + 2\sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{G^o} \hat{\alpha}_g^o f(Y_{i,obs};\hat{\zeta}_g^o) \right\} - \log n\phi(G^o) > 0,$$

in probability, as $n \to \infty$. Similarly, we can conclude that G^o is the minimizer of (5.17), which completes proof of *Step 2*.

5.10 Appendix B: More simulation results

Table 5.5: Simulation results for the simulation study II from 2,000 Monte Carlo studies. The numbers we presented are average coverage probabilities and interval lengths of 95% confidence intervals (×100).

Model	Method	Response	θ_2	θ_3	θ_4	P_2	P_3	P_4
M1	Full	MCAR	60.8(94.8)	60.8(94.9)	60.8(94.7)	8.3(94.7)	8.3(94.5)	8.3(94.3)
	CC		84.9(94.7)	84.9(94.2)	84.9(94.0)	11.7(94.8)	11.7(94.9)	11.7(95.1)
	MICE		61.1(94.8)	61.0(94.7)	61.2(94.7)	8.5(94.7)	8.5(94.7)	8.5(94.4)
	FIGURE		63.1(95.3)	62.6(95.0)	63.0(95.0)	8.7(94.8)	8.6(95.3)	8.7(94.6)
	Full	MAR	60.8(95.0)	60.8(95.2)	60.8(95.3)	8.8(96.0)	8.8(95.2)	8.8(95.7)
	CC		83.1(42.4)	83.1(42.8)	83.0(41.9)	12.1(52.8)	12.1(52.5)	12.1(52.3)
	MICE		61.3(95.0)	61.2(95.2)	61.2(95.5)	9.0(95.9)	9.0(95.2)	9.0(95.0)
	FIGURE		63.6(95.2)	62.9(95.2)	62.8(95.3)	9.2(95.5)	9.2(95.7)	9.2(95.8)
M2	Full	MCAR	60.8(95.0)	60.8(94.5)	60.8(94.8)	8.8(94.3)	8.8(94.8)	8.8(94.3)
	CC		84.9(94.5)	84.9(94.8)	84.9(94.9)	12.3(94.0)	12.3(94.3)	12.3(94.0)
	MICE		61.3(94.8)	61.2(94.0)	61.2(94.2)	9.0(94.0)	9.0(94.5)	9.0(94.2)
	FIGURE		63.5(95.0)	62.7(94.7)	62.7(94.8)	9.2(94.2)	9.2(94.6)	9.2(94.5)
	Full	MAR	60.8(95.0)	60.8(95.2)	60.8(95.3)	8.8(96.0)	8.8(95.2)	8.8(95.7)
	CC		83.1(42.4)	83.1(42.8)	83.0(41.9)	12.1(52.8)	12.1(52.5)	12.1(52.3)
	MICE		61.3(95.0)	61.2(95.2)	61.2(95.5)	9.0(95.9)	9.0(95.2)	9.0(95.0)
	FIGURE		63.6(95.2)	62.9(95.2)	62.8(95.3)	9.2(95.5)	9.2(95.7)	9.2(95.8)
M3	Full	MCAR	17.5(95.0)	17.5(96.2)	17.5(94.5)	7.9(94.8)	7.9(94.3)	7.9(95.4)
	CC		24.4(95.1)	24.4(95.2)	24.4(95.4)	11.0(95.2)	11.0(93.9)	11.0(95.7)
	MICE		19.3(94.5)	19.2(95.2)	19.4(95.0)	8.7(94.2)	8.7(94.5)	8.7(95.5)
	FIGURE		19.8(95.2)	19.6(96.0)	19.9(95.3)	8.7(95.2)	8.6(94.8)	8.7(95.9)
	Full	MAR	17.5(95.0)	17.5(93.9)	17.5(95.5)	7.9(94.6)	7.9(94.8)	7.9(94.8)
	CC		24.4(95.3)	24.5(93.8)	24.4(95.2)	11.0(94.6)	11.0(94.2)	11.0(94.5)
	MICE		19.4(95.0)	19.3(94.7)	19.5(95.5)	8.8(93.8)	8.7(94.3)	8.8(94.4)
	FIGURE		19.9(95.5)	19.8(95.2)	19.9(96.0)	8.7(95.5)	8.6(94.8)	8.7(96.2)

5.11 Appendix C: Proof of Lemma 5.1

Bacharoglou (2010) established the following lemma.

Lemma 5.3. *For every density function f_0 of random variable Y and every $\epsilon > 0$, there exist normal distributions $\phi_1, \phi_2, \cdots, \phi_G$ and positive numbers $\alpha_1, \cdots, \alpha_G$ with $\sum_g^G \alpha_g = 1$, such that*

$$\sup |f_0 - \sum_{g=1}^{G} \alpha_g \phi_G| < \epsilon,$$
$$\|f_0 - \sum_{g=1}^{G} \alpha_g \phi_G\|_1 < \epsilon.$$

Now, assume that $f = \sum_{g=1}^{G} \alpha_g \phi_G$. Then, we can establish the following lemma.

Lemma 5.4. *There exist $G = G(\epsilon)$, such that*

$$\|f_0 - \hat{f}\|_1 < \epsilon,$$
$$\sup |f_0 - \hat{f}| < \epsilon,$$

with probability one, where $\hat{f} = \sum_{g=1}^{G} \hat{\alpha}_g \hat{\phi}_g$ is a minimizer of

$$-\frac{1}{n} \sum_{i=1}^{n} \log f(Y_i)$$

and f is obtained by minimizing

$$E_0 \left(\log \frac{f_0}{\sum_{g=1}^{G} \alpha_g \phi_G} \right). \tag{5.35}$$

Proof. Note that the minimizer of (5.35) is unique with probability one under assumption $\alpha_1 \geq \alpha_2 \geq \ldots \alpha_G$. Moreover, if f_1 satisfies Lemma 5.3, then

$$E_0(\log \frac{f_0}{f_1}) \leq E_0 \left\{ \log f_0 - \log(f_0 - \epsilon) \right\} \leq \epsilon,$$

for any $\epsilon > 0$. Thus,

$$E_0 \left(\log f_0 - \log f \right) \geq E_0 \left(\log f_0 - \log f_1 \right).$$

Therefore, f_1 is a minimizer and $f = f_1$ with probability one. If $\sup |f_0 - f| \geq \epsilon$ at Y_0, then there exist a ball $B_r(Y_0)$, such that $|f_0(Y) - f(Y)| \geq \epsilon$ for any $Y \in B_r(Y_0)$, since f_0 is continuous. Then,

we can obtain that $\|f_0 - f\|_1 \geq V\epsilon$, where V is the volume of $B_r(Y_0)$. This is a contradiction. Thus, we complete the proof of Lemma 5.4 for f. Now, consider $\hat{f} = \sum_{g=1}^{G} \hat{\alpha}_g \hat{\phi}_g$, where $\hat{f}$ is a minimizer of

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_0(Y_i)}{f(Y_i)} \to E_0 \left(\log f_0 - \log f\right),$$

with probability one, where $\{Y_1, \cdots, Y_n\}$ are n IID realizations from f_0. Thus, Lemma 5.4 holds for $\hat{f}$ with probability one. □

Using Lemma 5.4, let $G = \epsilon^{-\gamma} + G_0$, where γ depends on f_0. If f_0 is a Gaussian mixture model, then $\gamma = 0$. Otherwise, $\gamma > 0$. Without loss of generality, we assume $\Sigma = \mathrm{I}$. Thus, the parameters are $\zeta_g = (\mu_g, \alpha_g)$. Let $\zeta = (\zeta_1, \cdots, \zeta_G)$.

$$\mathrm{var}\left\{\hat{f} - f_0\right\} \cong \mathrm{var}\left\{f - f_0 + \frac{\partial f}{\partial \zeta^T}(\hat{\zeta} - \zeta)\right\}.$$

Thus,

$$\mathrm{var}\left\{\hat{f} - f_0\right\} \cong E\left(\frac{\partial f}{\partial \zeta^T}\right) \mathrm{var}\left(\hat{\zeta} - \zeta\right) E\left(\frac{\partial f}{\partial \zeta}\right). \tag{5.36}$$

Under MAR assumption, we have

$$\mathrm{var}\left(\hat{\zeta} - \zeta\right) \cong \frac{\mathrm{I}_{obs}^{-1}}{n},$$

where I_{obs} is the fisher information matrix from the observed likelihood.

Since $f = \sum_{g=1}^{G} \alpha_g \phi_g$, we have

$$\frac{\partial^2 \log f}{\partial \mu_g \partial \mu_g^T} = -\frac{\alpha_g \phi_g}{f} \mathrm{I}_{p\times p} - \frac{\alpha_g^2 \phi_g^2}{f^2}(Y - \mu_g)(Y - \mu_g)^T.$$

Now, consider the divergence case in the sense that $G \to \infty$, as $n \to \infty$. Since we assume that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_G$, we have $\alpha_G \leq O(1/G)$. $G = \epsilon^{-\gamma} + G_0$. Then $\alpha_G \leq O(\epsilon^{\gamma})$. Note that

$$\mathrm{I}_{obs} = -E\left(\frac{\partial^2 f_{obs}}{\partial \zeta \partial \zeta^T}\right),$$

where f_{obs} is the marginal density of f corresponding to the observed part. We can decompose I_{obs} as

$$\mathrm{I}_{obs} = \begin{pmatrix} A & B \\ B^T & D \end{pmatrix},$$

where $D = E_0\left\{\frac{\alpha_G\phi_G}{f}\mathrm{I}_{p\times p} + \frac{\alpha_G^2\phi_G^2}{f^2}(Y-\mu_G)(Y-\mu_G)^T\right\}$ Then, applying the block inverse form, we can obtain that

$$\mathrm{I}_{obs}^{-1} = \begin{pmatrix} \star & \star \\ \star & (D - CA^{-1}B)^{-1} \end{pmatrix} \tag{5.37}$$

If we assume $E_0\left\{f^{-2}(Y)\right\} < \infty$, then $D = O(\alpha_G)$. Thus, $\mathrm{I}_{obs}^{-1} \cong O(\alpha_G^{-1}) = O(\epsilon^{-\gamma})$

Thus, we can summarize the approximation of GMM as

$$\sup_Y |\hat{f} - f_0| = O(\epsilon),$$
$$\mathrm{var}(\hat{f} - f_0) = O(\epsilon^{-\gamma}n^{-1}),$$

if $E\left(\partial f/\partial\zeta^T\right)$ are bounded. This assumption is true for GMM.

5.12 Appendix D: Proof of Theorem 5.2

In this section, we will show the $\sqrt{n}$-consistency of the proposed estimator $\hat{\theta}_{FIGURE}$. Note that, $\hat{\theta}_{FIGURE}$ is a solution of

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{M_g} w_{igj}^* U(\theta; Y_i^{*(gj)}) = 0..$$

Then, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{M_g} w_{igj}^* U(\theta; Y_i^{*(gj)}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[R_i U(\theta; Y_i) + (1-R_i)E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, \hat{f}\right\}\right] + O_p\left(\frac{1}{\sqrt{M}}\right),
\end{aligned}$$

where $M = \min_g\{M_g\}$ and we can let $M \to \infty$. Ignoring the smaller term, we can rewrite the estimating equation as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{M_g} w_{igj}^{*} U(\theta; Y_i^{*(gj)}) \\
&\cong \frac{1}{n}\sum_{i=1}^{n}\left[R_i U(\theta; Y_i) + (1-R_i) E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\}\right] \\
&+\frac{1}{n}\sum_{i=1}^{n}\left[(1-R_i) E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\}\right. \\
&\left.+(1-R_i) E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, \hat{f}\right\}\right]. \\
&= J_1 + J_2.
\end{aligned}$$

Note that,

$$J_1 = \frac{1}{n}\sum_{i=1}^{n}\left[R_i U(\theta; Y_i) + (1-R_i) E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\}\right]$$

is an unbiased estimating equation for θ.

For J_2 term, we have

$$\begin{aligned}
&\left\| E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\} - E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, \hat{f}\right\}\right\|_1 \\
&= \int \left| U(\theta; Y_{i,obs}, Y_{i,mis}) \left\{ \frac{\hat{f}(Y_i)}{\hat{f}_{obs}(Y_{i,obs})} - \frac{f_0(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\} \right| f_0(Y_i) dY_i \\
&\leq E_0 \left| U(\theta; Y_{i,obs}, Y_{i,mis}) \right| \left\| \frac{\hat{f}(Y_i)}{\hat{f}_{obs}(Y_{i,obs})} - \frac{f_0(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_1 .
\end{aligned}$$

Assume $E\left|U(\theta; Y_{i,obs}, Y_{i,mis})\right| < \infty$. Moreover,

$$\left\| \frac{\hat{f}(Y_i)}{\hat{f}_{obs}(Y_{i,obs})} - \frac{f_0(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_1 \leq \left\| \frac{\hat{f}(Y_i)}{\hat{f}_{obs}(Y_{i,obs})} - \frac{\hat{f}(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_1 + \left\| \frac{\hat{f}(Y_i)}{f_{0,obs}(Y_{i,obs})} - \frac{f_0(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_1 .$$

For the first term, we can show

$$\begin{aligned}
\left\| \frac{\hat{f}(Y_i)}{\hat{f}_{obs}(Y_{i,obs})} - \frac{\hat{f}(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_1 &\leq \|\hat{f}(Y)\|_1 \|\hat{f}_{obs}(Y_{obs}) - f_{0,obs}(Y_{i,obs})\|_1 \frac{1}{f_{0,obs}^2(Y_{i,obs})(1-\epsilon)} \\
&\leq \frac{\|\hat{f}(Y)\|_1}{f_{0,obs}^2(Y_{i,obs})(1-\epsilon)} \|\hat{f} - f\|_1 . \\
&\leq C_3 \epsilon. \qquad (5.38)
\end{aligned}$$

For the second term, we have

$$\left\| \frac{\hat{f}(Y_i)}{f_{0,obs}(Y_{i,obs})} - \frac{f_0(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_1 = \frac{1}{f_{0,obs}(Y_{i,obs})} \|\hat{f} - f_0\|_1 \leq C_4 \epsilon. \tag{5.39}$$

Using (5.38) and (5.39), we can conclude that

$$\left\| E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\} - E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, \hat{f}\right\} \right\|_1 = O(\epsilon). \tag{5.40}$$

Next, we can show the variance

$$\begin{aligned} &\text{var}\left[E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\} - E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, \hat{f}\right\}\right] \\ &\leq E\left| E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\} - E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, \hat{f}\right\} \right|^2. \end{aligned}$$

Using the similar technique above, we can show that

$$\begin{aligned} &\text{var}\left[E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, f_0\right\} - E\left\{U(\theta; Y_{i,obs}, Y_{i,mis}) \mid Y_{i,obs}, \hat{f}\right\}\right] \\ &\leq E\left|U(\theta; Y_{i,obs}, Y_{i,mis})\right|^2 \left\| \frac{\hat{f}(Y_i)}{\hat{f}_{obs}(Y_{i,obs})} - \frac{f_0(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_2^2. \end{aligned}$$

Assume $E\left|U(\theta; Y_{i,obs}, Y_{i,mis})\right|^2 < \infty$. Moreover,

$$\left\| \frac{\hat{f}(Y_i)}{\hat{f}_{obs}(Y_{i,obs})} - \frac{f_0(Y_i)}{f_{0,obs}(Y_{i,obs})} \right\|_2^2 \leq C_5 \|\hat{f} - f_0\|_2^2 = C_5 \left\{ \text{var}(\hat{f} - f_0) + \|\hat{f} - f_0\|_1^2 \right\} = O(\epsilon^2 + \epsilon^{-\gamma} n^{-1}) \tag{5.41}$$

Using (5.40) and (5.41), we can prove that

$$\begin{aligned} J_2 &= \frac{1}{n} \sum_{i=1}^{n} (1 - R_i) O(\epsilon) + O_p\left[\sqrt{\frac{1}{n^2} \sum_{i=1}^{n} \pi_i (1 - \pi_i) O(\epsilon^2 + \epsilon^{-\gamma} n^{-1})} \right] \\ &= O_p(\epsilon) + O_p\left\{ \left(\epsilon^2 n^{-1} + \epsilon^{-\gamma} n^{-2} \right)^{1/2} \right\}, \end{aligned}$$

where $\pi_i = \Pr(R_i = 1 \mid Y_i)$.

If $\epsilon = O(n^{-1/(2-\Delta)})$ with $\Delta \in (0, 2)$ and $\gamma \in (0, 2)$, we have

$$J_2 = o_p(n^{-1/2}). \tag{5.42}$$

Finally, using (5.42), we have

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{g=1}^{G} \sum_{j=1}^{M_g} w_{igj}^* U(\theta; Y_i^{*(gj)}) \cong J_1 + o_p(n^{-1/2}).$$

Bibliography

Bacharoglou, A. (2010). Approximation of probability distributions by convex mixtures of gaussian measures. *Proceedings of the American Mathematical Society*, 138(7):2619–2628.

Berg, E., Kim, J.-K., and Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4(4):436–462.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Chen, H. Y. (2010). Compatibility of conditionally specified models. *Statistics & Probability Letters*, 80(7-8):670–677.

Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103(484):1674–1683.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):81–87.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, 39(1):1–38.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

Huang, T., Peng, H., and Zhang, K. (2017). Model selection for gaussian mixture models. *Statistica Sinica*, 27(1):147–169.

Im, J., Kim, J.-k., and Fuller, W. A. (2018). Two-phase sampling approach to fractional hot deck imputation. *Unpublished Manuscript*.

Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Haung, W.-C. (2007). Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation. pages 3211–3218.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1):119–132.

Kim, J. K., Michael Brick, J., Fuller, W. A., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):509–521.

Kim, J. K. and Shao, J. (2013). *Statistical methods for handling incomplete data*. CRC Press.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages 1–163. JSTOR.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied statistics*, 36(3):318–324.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558.

Oliver, J. J., Baxter, R. A., and Wallace, C. S. (1996). Unsupervised learning using mml. In *ICML*, pages 364–372.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Wallace, C. S. and Dowe, D. L. (1999). Minimum message length and kolmogorov complexity. *The Computer Journal*, 42(4):270–283.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 37(1):490–517.

Windham, M. P. and Cutler, A. (1992). Information ratios for validating mixture analyses. *Journal of the American Statistical Association*, 87(420):1188–1192.

Yang, S. and Kim, J. K. (2016a). Fractional imputation in survey sampling: A comparative review. *Statistical Science*, 31(3):415–432.

Yang, S. and Kim, J. K. (2016b). A note on multiple imputation for method of moments estimation. *Biometrika*, 103(1):244–251.

Table 5.1: Simulation results for the simulation study I from 2,000 Monte Carlo studies. The numbers we presented are RMSE in (5.32).

Model	Method	Response	θ_2	θ_3	θ_4	P_2	P_3	P_4
M1	Full	MCAR	100	100	100	100	100	100
	CC		195	195	195	193	190	195
	MICE		101	101	101	100	100	101
	FIGURE		101	101	102	100	100	101
	Full	MAR	100	100	100	100	100	100
	CC		9529	9359	9138	6517	6266	6439
	MICE		100	101	101	99	98	100
	FIGURE		100	100	102	99	99	103
M2	Full	MCAR	100	100	100	100	100	100
	CC		187	186	188	185	187	183
	MICE		104	105	105	102	102	104
	FIGURE		103	103	103	99	99	99
	Full	MAR	100	100	100	100	100	100
	CC		8436	8305	8100	6303	6104	6204
	MICE		108	108	109	98	99	101
	FIGURE		107	107	106	98	100	100
M3	Full	MCAR	100	100	100	100	100	100
	CC		207	199	200	196	196	190
	MICE		108	111	114	118	114	130
	FIGURE		107	109	112	112	110	126
	Full	MAR	100	100	100	100	100	100
	CC		1251	389	183	152	934	223
	MICE		104	104	101	122	240	152
	FIGURE		104	105	99	110	183	144
M4	Full	MCAR	100	100	100	100	100	100
	CC		195	199	197	195	196	193
	MICE		103	102	103	207	171	157
	FIGURE		107	105	105	144	141	162
	Full	MAR	100	100	100	100	100	100
	CC		1251	1147	866	735	1093	890
	MICE		104	103	104	227	189	265
	FIGURE		104	105	103	147	145	226

Table 5.2: Simulation results for the simulation study II from 2,000 Monte Carlo studies. The numbers we presented are RMSE in (5.32) and coverage probabilities of 95% confidence intervals.

Model	Method	Response	θ_2	θ_3	θ_4	P_2	P_3	P_4
M1	Full	MCAR	100	100	100	100	100	100
	CC		196	198	199	197	202	197
	MICE		102	101	102	103	103	104
	CFIGURE		105	103	104	103	101	102
	Full	MAR	100	100	100	100	100	100
	CC		1108	1106	1103	922	918	904
	MICE		102	102	102	109	106	107
	CFIGURE		107	104	104	105	101	102
M2	Full	MCAR	100	100	100	100	100	100
	CC		196	199	198	191	200	193
	MICE		102	102	102	107	107	107
	CFIGURE		104	101	102	103	102	102
	Full	MAR	100	100	100	100	100	100
	CC		1108	1106	1103	922	918	904
	MICE		102	102	102	109	106	107
	CFIGURE		107	104	104	105	101	102
M3	Full	MCAR	100	100	100	100	100	100
	CC		188	203	190	184	197	183
	MICE		121	123	122	127	125	120
	CFIGURE		118	117	116	115	112	111
	Full	MAR	100	100	100	100	100	100
	CC		186	199	196	195	190	197
	MICE		120	119	121	126	122	126
	CFIGURE		115	113	117	112	106	110

Table 5.3: Imputation results for the monthly retail trade survey. Parameter estimation and 95% confidence lower and upper bounds.

parameter	FIGURE	lower bound	upper bound	Truth
Mean of Sales00 ($\times 10^{-6}$)	2.28	2.10	2.46	2.30
Skewness of Sales00	49.68	24.15	74.87	49.67
Mean of Inventories00 ($\times 10^{-6}$)	4.76	4.42	5.10	4.81
Skewness of Inventories00	40.00	10.28	69.40	39.02
Correlation of Sales00 and Inventories00	0.97	0.94	0.99	0.97

Table 5.4: Simulation results for the simulation study I from 2,000 Monte Carlo studies. The numbers we presented are average coverage probabilities and interval lengths of 95% confidence intervals (×100).

Model	Method	Response	θ_2	θ_3	θ_4	P_2	P_3	P_4
M1	Full	MCAR	60.9(94.8)	60.8(94.6)	60.8(94.5)	8.3(94.8)	8.3(94.5)	8.3(94.8)
	CC		84.9(94.3)	84.9(94.5)	84.9(94.6)	11.7(94.7)	11.7(95.0)	11.7(95.0)
	MICE		61.0(94.8)	61.0(94.7)	61.2(94.4)	8.4(95.4)	8.4(95.5)	8.5(95.5)
	FIGURE		60.9(94.2)	61.0(94.2)	61.5(94.5)	8.4(95.2)	8.4(94.8)	8.5(94.6)
	Full	MAR	60.8(94.5)	60.8(94.7)	60.8(95.2)	8.3(94.3)	8.3(94.8)	8.3(94.6)
	CC		72.7(0.0)	72.8(0.0)	72.9(0.0)	9.0(0.0)	9.0(0.0)	9.0(0.0)
	MICE		61.0(94.8)	61.1(94.6)	61.2(94.7)	8.5(95.5)	8.5(95.3)	8.5(95.8)
	FIGURE		61.1(93.9)	61.3(94.0)	61.9(94.7)	8.4(94.8)	8.5(94.7)	8.7(95.5)
M2	Full	MCAR	66.6(95.1)	66.7(95.2)	66.7(94.2)	8.3(94.0)	8.3(94.8)	8.3(94.3)
	CC		93.2(95.3)	93.2(95.2)	93.1(94.3)	11.7(94.4)	11.7(94.5)	11.7(95.0)
	MICE		67.9(95.2)	68.0(95.2)	68.0(94.2)	8.6(95.3)	8.6(95.0)	8.6(95.2)
	FIGURE		68.6(94.6)	69.1(94.8)	68.9(94.0)	8.5(95.0)	8.6(94.7)	8.6(95.2)
	Full	MAR	66.7(94.0)	66.6(94.3)	66.7(94.5)	8.3(94.0)	8.3(93.9)	8.3(93.9)
	CC		78.0(0.0)	78.1(0.0)	78.0(0.0)	9.0(0.0)	9.0(0.0)	9.0(0.0)
	MICE		67.7(93.5)	67.8(94.0)	67.8(93.2)	8.6(95.6)	8.6(95.5)	8.7(95.5)
	FIGURE		69.5(93.8)	69.6(94.0)	70.1(93.7)	8.5(94.9)	8.5(94.5)	8.6(95.2)
M3	Full	MCAR	24.8(94.7)	30.3(94.3)	24.7(94.2)	8.6(95.0)	8.7(95.2)	8.6(94.5)
	CC		34.5(93.8)	42.3(94.0)	34.6(94.2)	12.0(95.4)	12.1(95.2)	12.0(94.2)
	MICE		25.8(95.2)	32.0(94.2)	26.5(94.9)	9.2(95.0)	9.4(95.3)	9.4(93.7)
	FIGURE		26.3(95.0)	32.9(94.7)	27.2(94.8)	9.1(94.7)	9.3(94.7)	9.4(93.4)
	Full	MAR	24.8(95.0)	30.3(94.8)	24.8(94.8)	8.6(95.0)	8.7(95.3)	8.6(94.5)
	CC		36.9(61.1)	44.3(71.8)	36.8(61.2)	12.1(71.0)	12.2(79.4)	12.1(71.2)
	MICE		25.9(94.5)	32.1(94.6)	26.6(94.6)	9.3(94.4)	9.4(95.9)	9.5(92.5)
	FIGURE		26.3(94.6)	32.9(94.8)	27.1(95.2)	9.1(93.9)	9.3(94.3)	9.5(93.2)
M4	Full	MCAR	42.7(94.7)	133.1(94.8)	429.5(93.2)	8.5(94.9)	8.7(94.7)	8.8(95.4)
	CC		60.0(94.2)	188.3(94.7)	611.4(93.4)	11.8(94.4)	12.2(94.0)	12.2(95.3)
	MICE		43.2(94.6)	133.4(94.3)	431.5(93.0)	8.9(84.5)	9.1(88.8)	9.4(91.6)
	FIGURE		46.5(95.0)	135.2(93.9)	436.4(91.5)	9.6(93.1)	9.3(91.7)	9.5(90.3)
	Full	MAR	42.8(95.1)	134.3(94.9)	438.2(93.1)	8.5(95.4)	8.7(95.6)	8.8(95.4)
	CC		65.5(45.7)	212.3(55.4)	714.6(74.8)	11.4(57.4)	12.2(44.2)	12.0(52.9)
	MICE		43.4(95.8)	135.5(95.1)	447.3(93.5)	9.0(81.5)	9.1(87.9)	9.4(80.4)
	FIGURE		45.0 (95.2)	136.2(94.2)	444.0(93.0)	9.9(93.4)	9.2(92.4)	9.8(86.1)

CHAPTER 6. SUMMARY AND CONCLUSION

This dissertation investigates four topics in missing data: Bayesian propensity score estimation, Sparse propensity score estimation, a profile approach to semiprametric estimation with nonignorable nonresponse and semiparametric fractional imputation for handling multivariate missing data.

In Chapter 2, we propose a new Bayesian inference using the idea of approximate Bayesian computation. The proposed Bayesian method is further extended to incorporate auxiliary information from full sample. The proposed method can be widely applicable to causal inference and combing information from different sources. In Chapter 3, Bayesian approach to propensity score estimation using the Spike-and-Slab prior for the response propensity model is proposed. Extension of our proposed method to nonignorable nonresponse is a topic for future research. In Chapter 4, we propose a semiparametric method using the maximum profile likelihood to achieve robust estimation under nonignorable nonresponse. Then, we also propose a test procedure to check if the response mechanism is missing at random. The bootstrap procedures are developed to compute the empirical distribution of the proposed test statistic. In Chapter 5, a unified fractional imputation method using Gaussian mixture models is proposed. The proposed method automatically selects mixture components using Bayesian information criterion. The flexible model assumption and efficient computation are main advantages of the proposed method. R software packages for the proposed methods in this thesis are under development.